Appendix: Supplementary Material for Paper titled "Software Fairness: An Analysis and Survey"

EZEKIEL SOREMEKUN, Singapore University of Technology and Design, Singapore
MIKE PAPADAKIS, MAXIME CORDY, and YVES LE TRAON, SnT, University of Luxembourg, Luxembourg

Summary

This document is the appendix (or supplementary material) for the paper titled "Software Fairness: An Analysis and Survey". It provides additional or more detailed tables and figures to support the reported findings in the original paper.

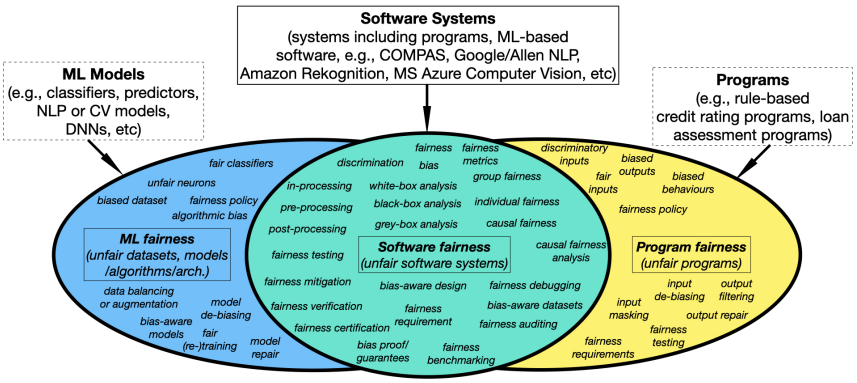


Fig. 1. Taxonomy and interplay of ML fairness and Software Fairness

Authors' addresses: Ezekiel Soremekun, ezekiel_soremekun@sutd.edu.sg, Singapore University of Technology and Design, 8 Somapah Rd, Singapore, Singapore, Singapore, 487372; Mike Papadakis, michail.papadakis@uni.lu; Maxime Cordy, maxime.cordy@uni.lu; Yves Le Traon, Yves.LeTraon@uni.lu, SnT, University of Luxembourg, 6, rue Richard Coudenhove-Kalergi, Luxembourg, Luxembourg, Luxembourg, L-1359.

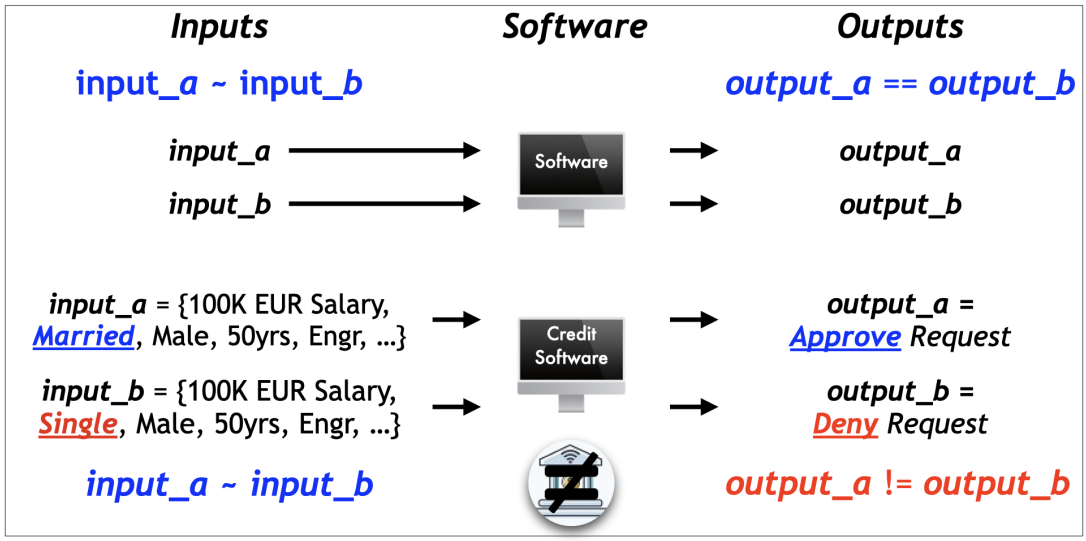


Fig. 2. Program fairness property showing fair program behavior (`output_a == output_b`) and unfair program behavior (`output_a != output_b`) using a (credit approval) software

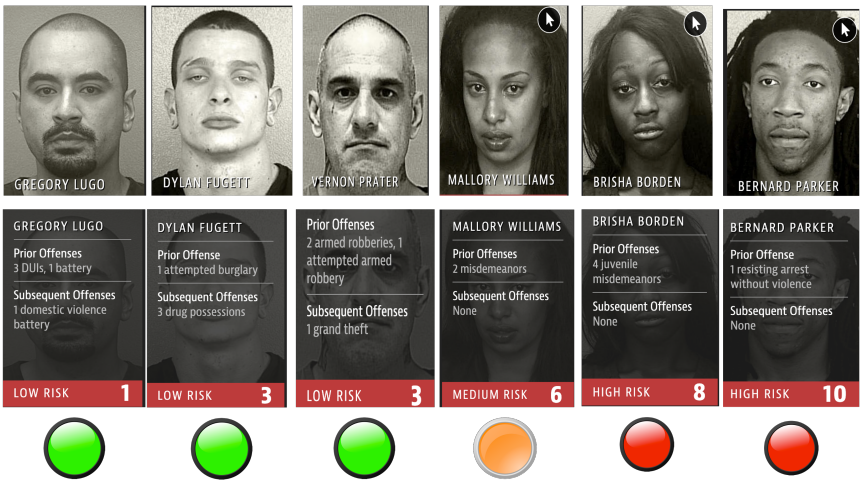


Fig. 3. An illustrative example of real-world discrimination (unfair software behavior) in the COMPAS software, a program that predicts recidivism – the likelihood of committing a future crime (adapted from Angwin et al. [12, 13])

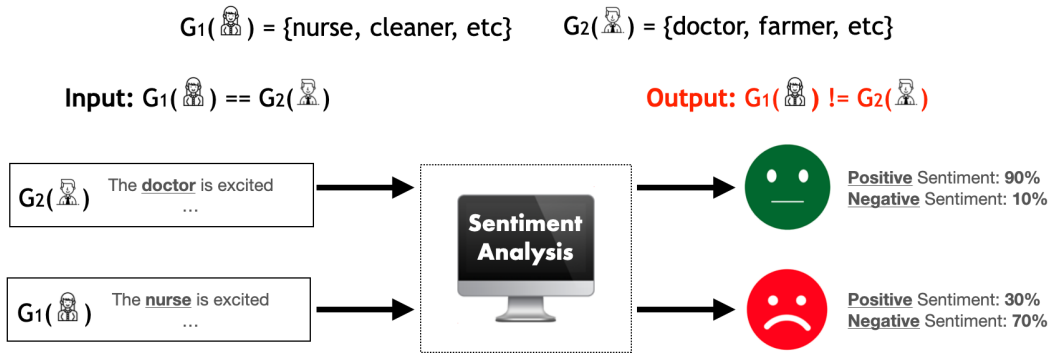


Fig. 4. Example of Group Fairness Violation using a Sentiment Analysis AI System

Table 1. Excerpt of Publication Details ("#" means "number of")

Domain (#Pubs.)	**Type**	**Venue**		**Publications**		
		#Venues	**Sample Venue(s)**	**#Pubs**	**Example Publications**	**Years**
Software Engineering (SE), Programming Languages (PL) & Security	*Conference*	9	ASE, CAV, EuroS&P, FSE, GECCO, ICSE, OOPSLA, TrustCom, ISSTA	28	[24, 42, 50, 61, 62, 78, 101, 141]	2017-21
	Journal	4	EMSE, JSS, RE, TSE	7	[14, 20, 21, 57, 136, 137, 164]	2009-21
	Other	1	ICSE-C	1	[6]	2021
Natural language processing (NLP)	*Conference*	2	ACL, EMNLP	6	[28, 29, 56, 123, 129, 166]	2017-21
Artificial Intelligence (AI) & Machine Learning (ML)	*Conference*	8	AAAI, AISTATS, ICML, NeurIPS, PMLR	17	[4, 96, 109, 158] [3, 35, 83, 92, 121, 154]	2013-21
	Other	1	HRLC	1	[165]	2021
Computer Vision (CV)	*Conference*	2	ICCV, CVPR	3	[94, 150, 151]	2019-20
	Workshop	1	ECCV	1	[157]	2020
Fairness-targets	*Conference*	2	AAAI-AIES, Facct	21	[9, 22, 27, 36, 58, 73, 93, 125, 134]	2017-21
	Workshop	2	FairWare, FATE	8	[15, 30, 46, 47, 80, 148, 153]	2018-19
Big Data, Data Mining (DM), & Knowledge Discovery (KD)	*Conference*	8	DMKD, ECML-PKDD, EDBT, KDD, ICDM, ICEDT, ICMD, LAK	18	[34, 89, 99, 169] [43, 52, 88, 91, 139]	2010-21
	Journal	5	Big Data, Inf. Science, JDIQ, KAIS, SIGMOD-Record	5	[1, 48, 87] [90, 128]	2012-19
	Workshop	2	BSDUC, KDD-XAI	2	[69, 124]	2018-19
Human Factors & Usability	*Conference*	1	CHI	3	[44, 75, 98]	2019-21
	Journal	1	IWC	1	[32]	2016
Others	*Conference*	3	CCCT, VAST, WWW	6	[33, 51, 70, 86, 97]	2009-20
	Journal	6	CACM, DGRP, Scientific-data, SSRN, IBM Journal of R & D	10	[19, 64, 82, 130, 131]	2018-21
	Workshop	1	CEUR Workshop	1	[138]	2019
	Other	3	arXiv, HRDAG, MS Tech. Report	15	[23, 45, 76, 84, 104, 127, 133]	2019-21

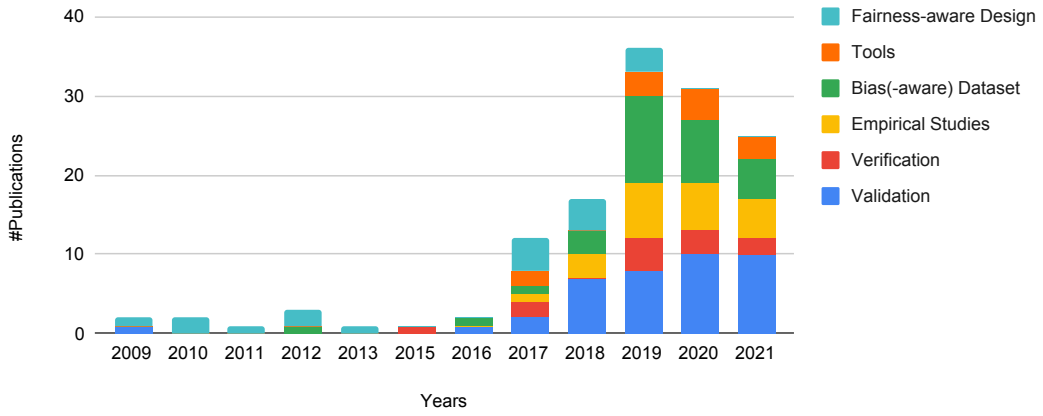
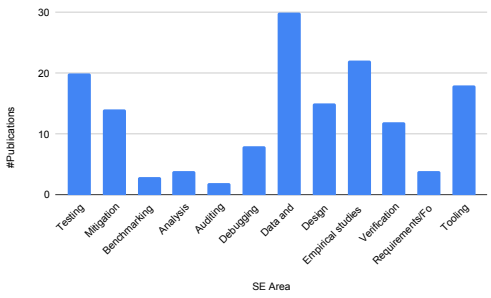


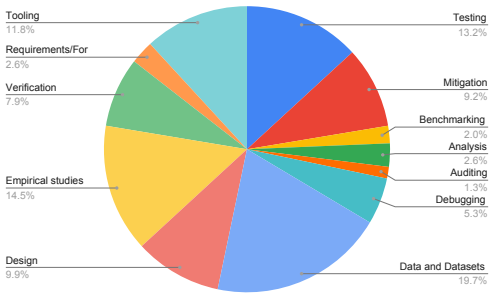
Fig. 5. Detailed Publication Trend by Year

Table 2. Purpose of Software Fairness Analysis (recent (2022 and 2023) publications are in bracket).

Categories	Sub-category	Description	#Pubs	Sample Works
Validation	Testing	generating discriminatory test inputs to expose fairness violations	20 (13)	[160, 163] ([10, 39, 100, 155])
	Mitigation	mitigating bias in software systems, e.g., via repair and prevention	14 (13)	[8] ([40, 62, 113, 118, 142, 161])
	Debugging	diagnosis and explanation of fairness violations	8 (3)	[42, 101, 137] ([110, 116]
	Auditing	analysing and measuring bias in software systems	2 (1)	[33, 96] ([162])
Verification	Verifiers	verifying that a system fulfills a fairness metric or goal	12 (1)	[7, 18, 65, 83] ([26])
	Certification	certifying that a system fulfills a fairness goal	4 (1)	[52, 132] ([26])
	Proof or Guarantees	providing a formal proof that a system achieves a fairness goal	2 (2)	[36, 109] ([66, 74])
Design	Requirements	requirement engineering and formalization of fairness properties	4 (2)	[21, 57, 103] ([17, 120])
	Bias-aware Design	designing fair systems and bias-aware software	15 (5)	[34, 80, 88] ([63, 81, 143])
Empirical Evaluation	Analysis	empirical studies about fairness concerns	22 (14)	[25, 30, 47, 166] ([72, 112, 152])
	Benchmarking	providing fair benchmarks or benchmarks for fairness evaluations	4 (3)	[24, 29, 78, 151] ([67, 77, 149])
Datasets	Bias in Datasets	studying biases in training and evaluation datasets	30	[35, 64, 150, 157]
	Bias-aware Datasets	developing unbiased or bias-aware datasets for better evaluation	5 (1)	[30, 124, 129, 130] ([149])
Tooling	Automatic	providing fully automatic tools for fairness analysis	18	[7, 19, 23, 65]
	Semi-automatic	building tools that require human interaction for fairness analysis	2	[33, 102]

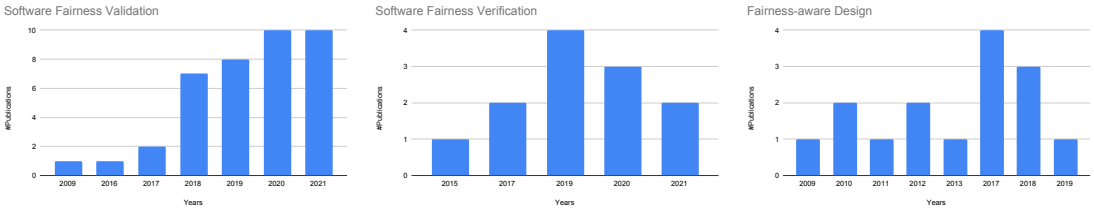


(a) Number of Software Fairness works in the SE community



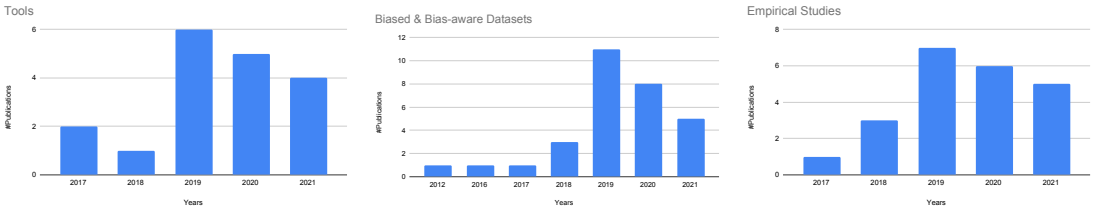
(b) Distribution of Software Fairness Works in the SE community

Fig. 6. Purpose of Fairness Analysis in SE community



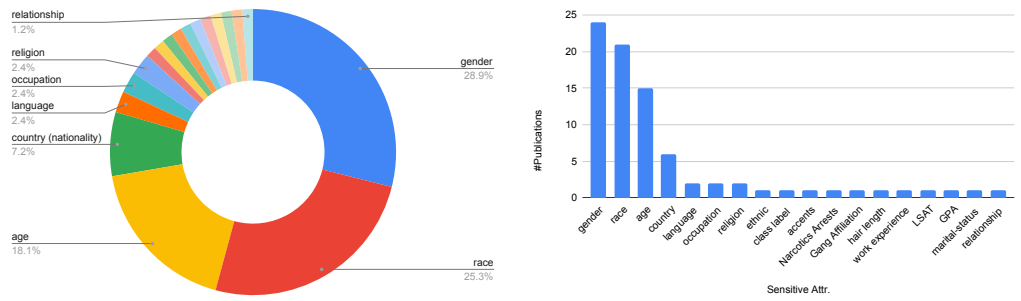
(a) Trends in Fairness Validation (b) Trends in Fairness Verification (c) Trends in Fairness-aware design

Fig. 7. Details of trends in Fairness Verification, Validation and Design



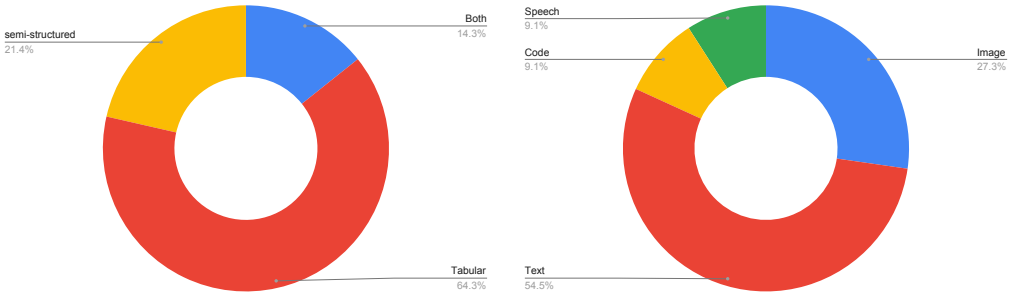
(a) Trends in Fairness Tools (b) Trends in Bias-related Datasets (c) Trends in Empirical Studies

Fig. 8. Details of trends in Fairness Tooling, Datasets and Empirical Studies



(a) Distribution of studied Sensitive Attributes (b) Number of Publications per Sensitive Attribute

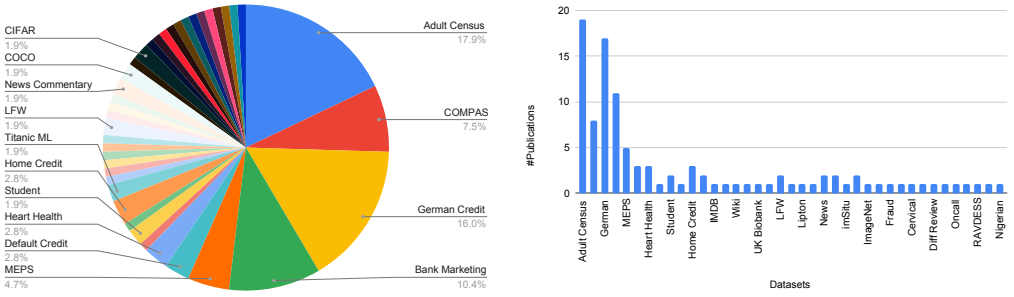
Fig. 9. Details of Biases, i.e., Sensitive/Protected attributes



(a) Proportion of Tabular or Semi-structured Datasets

(b) Distribution of Semi-structured Datasets

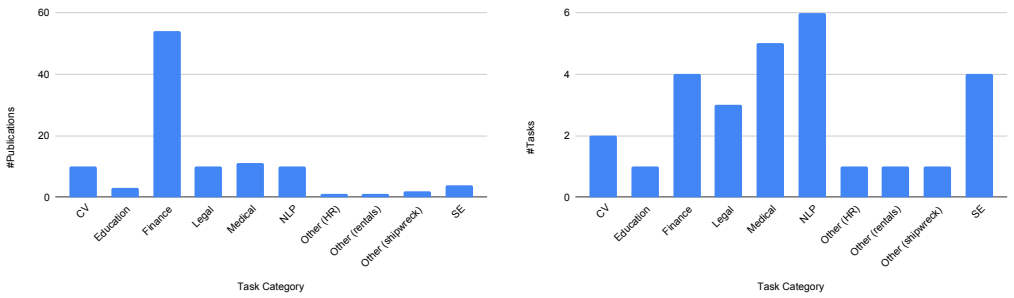
Fig. 10. Type of Studied Datasets



(a) Distribution of examined datasets

(b) Number of works (pubs) evaluating a specific dataset

Fig. 11. Details of examined Datasets



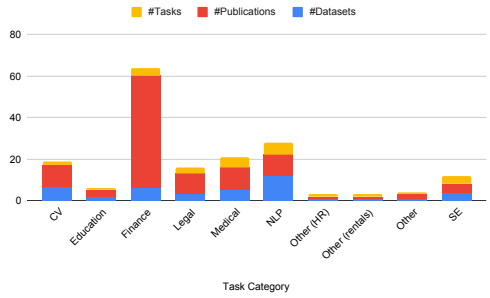
(a) Number of works (pubs) examining a task type

(b) Number of tasks studied in each category

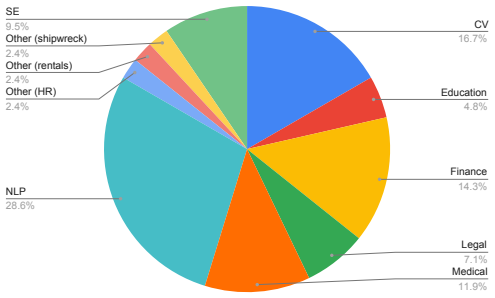
Fig. 12. Details of Task Categories

Table 3. Details of Tasks and Datasets employed in Software Fairness Analysis

Type	Task Category	#Data.	#Pubs	#Tasks	Tasks (Example Pubs)	Datasets (#Pubs)
Tabular	Education	2	3	1	academic performance [27]	Law School (1)
	Finance	6	54	4	income [5, 159, 163],	Adult Census (19)
					credit default [11, 25, 38]	German Credit (17), Default Credit (3), Home Credit (3)
					potential buyers [25, 37, 101]	Bank Marketing (11)
					fraud [5]	Fraud Detection (1)
	Legal	3	10	3	recidivism [33, 38, 78]	COMPAS (8)
					arrests [27]	Chicago Strategic Subject List (SSL) (1)
					US executions [5]	US Executions (1)
	Medical	5	11	5	medical expenditure [37, 101, 167]	MEPS (5)
					heart disease [37, 38]	Heart Health (3)
					heart failure [42]	Heart Failure (1)
					cancer risk [42]	Cervical Cancer (1)
	Other (HR)	1	1	1	hiring [27]	Lipton (1)
	Other (rentals)	1	1	1	car rentals [5]	Raw Car Rentals (1)
	Other (shipwreck)	1	2	1	shipwreck survival	Titanic ML (2)
Semi-structured	CV (image)	7	10	2	face detection	ClbA-IN (1), PPB (1), LFW (2)
					image recognition	COCO (2), imSitu (1), CIFAR (2), ImageNet (1)
	NLP (text, speech)	12	10	6	SA [14], CoRef [137], MLM [137]	Twitter (1), IMDB (1), EEC Dataset (1), Labor statistics (1)
					toxicity	Wiki Comment (1), Jigsaw Comments (1)
					machine translation	News Commentary (2)
					ASR [122]	Speech Accent Archive (1), RAVDESS (1), Multi speaker Corpora of the English Accents in the British Isles (1), Nigerian English speech dataset (1)
	SE (code)	4	4	4	Programs [42]	Bug2Commit (1), Diff Review (1), Code AutoComplete (1), Oncall Recommendation (1)



(a) Distribution of examined datasets



(b) Distribution of the Number of tasks per Task Category

Fig. 13. Details of Publications and Datasets for each Task category

Table 4. Excerpt of works performing Software Fairness Analysis ("Acc." means "level of software access")

Acc.	Approach	Goal	Problem	Main Idea	Core Technique
Black-box	LTDD [101]	Debugging	identifying biased features in training data	debugging biased features to build fair ML software.	data debugging, linear-regression
	Multiacc. Boost [96]	Auditing, Analysis, Mitigation	audit/mitigate multiaccuracy, i.e., group fairness for all subgroups	perform multiaccuracy audit and post-process models to achieve it	multiaccuracy auditing, post-processing
	Cito et al. [42]	Debugging, Mitigation	debug and isolate the cause of mispredictions in ML models	characterize the data on which the model performs poorly	rule induction
	ASTR-AEA [137]	Debugging, Testing, Mitigation	performing fairness testing without existing datasets	discover and diagnose fairness violations in NLP software	grammar-based testing
	Fair-Way [38]	Debugging, Testing, Mitigation	detect and explain how ML model acquires bias from training data	identify how ground truth bias affects ML fairness	multi-objective optimization, pre/in -processing
	FairSM-OTE [37]	Debugging, Testing, Mitigation	finding biased labels in training data generation	remove biased labels and balance data using sensitive attribute	situation testing, data balancing
	Fair-Vis [33]	Auditing, Analysis	auditing and analysing group fairness in ML model	visual analytics for the discovery and audit of (sub)group fairness	visual analytics, domain knowledge
	Flip-test [27]	Testing, Mitigation	testing individual fairness – similar treatment of protected statuses	discover individual (un)fairness and its associated features	optimal transport, flipset, distr. sampling
	Aequ-itas [147]	Testing, Mitigation	validation of fairness for arbitrary ML models?	generating discriminatory inputs to uncover fairness violations	directed testing, probabilistic search
	Themis [11, 31, 61]	Formaliz., Testing	formalize software fairness testing for casual discovering of discrimination	measure causal discrimination in software to direct fairness testing	input schema, causal relationships
	Aeque-Vox [122]	Debugging, Testing	testing group fairness for Automatic Speech Recognition (ASR) systems	group fairness testing by simulating different environments	ML robustness, test simulation, fault localization
	ExpGA [50]	Testing	current individual fairness testing methods suffer poor efficiency, effectiveness, and model specificity	fairness testing by modifying feature values using explanation results and genetic algorithm (GA)	genetic algorithm, feature mutation, search based testing
	CGFT [111]	Testing	Uneven distribution of fairness tests and variations in execution results	leverage combinatorial testing to generate evenly-distributed test suites	combinatorial testing, input coverage
	SG [5]	Testing	detecting the presence of individual discrimination in ML models.	auto-generation of test inputs for detecting individual discrimination.	symbolic execution, local explainability
	Bias-Finder [14]	Testing	Bias testers for SA systems rely on small, short, predefined templates	discover biased predictions in SA systems via metamorphic testing.	template curation, NLP techniques, metamorphic testing
	Biswas and Rajan [24]	Mitigation	understanding fairness characteristics in ML models from practice	empirical evaluation of fairness and mitigations on real-world ML models	empirical study
	Fairea [78]	Mitigation	what is the SE trade-off between accuracy and fairness?	benchmarking and quantifying the fairness-accuracy trade-off achieved by bias mitigation methods	model behaviour mutation
White-box	ADF [163, 164]	Testing	searching individual discriminatory instances	generating discriminatory inputs violating individual fairness via ML	gradient computation and clustering
	Deep-Inspect [144]	Testing	detecting confusion and bias errors at class-level	expose confusion and bias errors in image classifiers	class property violations, robustness
	EIDIG [160]	Testing, Mitigation	how to detect and improve individual fairness of a model	generating test cases that violate individual fairness	gradient descent, global/local search
	Neuron-Fair [167]	Testing, Mitigation, Analysis	interpretability, performance, and generalizability in bias testing	identifying biased neurons, i.e., neurons that cause discrimination	neuron activation, adversarial attacks
	Fair-Neuron [62]	Mitigation, Analysis	balancing accuracy-fairness trade-off without additional model(s)	detect neurons with contradictory optimization directions, and achieve trade-off via selective dropout	joint-optimization, adversarial game
Grey	Tizpaz-Niari [145]	Debugging, Testing, Mitigation	explaining fairness impact of hyper-parameters	identify the effect of parameters on software fairness	search based testing, statistical debugging
	CAT/ TransRepair [140, 141]	Testing, Mitigation	detecting inconsistency in machine translation (MT)	detect inconsistency bugs without access to human oracles	mutation testing, metamorphic testing, language model (BERT)

Table 5. Excerpt of Fairness Analysis Tools

Tool (Paper)	**Goal**	**Addressed Problem**	**Process. Stage**	**Approach**	**Access**
Fairkit-learn [84, 85]	Fair learning, Analysis	how to reason about and determine the trade-off between model quality (accuracy) and fairness	pre, & post	model search, visualisation	Grey
AIF360 [19]	Fair learning, Analysis	understanding how, when and why to use different bias handling algorithms in the model life-cycle	pre, in, & post	extensible architecture for analysing fairness metrics	Grey
POF [22]	Fair learning, Analysis	how to compute the "Pareto curve" of the trade-off between accuracy and fairness in the *regression* settings (*continuous* prediction/targeted values)	pre	fairness regularizers, Price of Fairness (PoF) metric	Grey
AITEST [6]	Testing	how to detect the presence of individual discrimination in ML models	post	symbolic execution and local explainability	Black
2AFC [102]	Testing	how to relate unobservable phenomena deep inside models with observable, outside quantities that we can measure from inputs and outputs	post	Test Experiments, Experimental Psychology, Psycophysics, two-alternative forced choice (2AFC)	Black
Pc-fair-ness [154]	Formalization, Analysis	how to bound path-specific counterfactual fairness, address their *identifiability*, i.e., whether they can be uniquely measured from observational data	post	parameterized causal modelling, linear programming, response-function variables, constraints	Black
BiasRV [156]	Testing	how to monitor and uncover biased predictions at runtime	post	automatic template generation, mutation, metamorphic relations	Black
Themis [11]	Formalization, Testing	how to formally define and test software fairness using a causality-based measure of discrimination	post	causal inference, schema -based test generation	Black
Fair-Square [7]	Formalization, Verification, Certification	how to verify or certify that a program meets a given fairness property	post	probabilistic reasoning, SMT solving, symbolic weighted -volume-computation algorithm	Black
FAT Fore-nsics [136]	Analysis, Auditing, certifying	how to inspect datasets (features), models and their prediction for fairness metrics	pre, in, & post	an inter-operable Python framework for fairness (FAT) algorithms	White, Black, & Grey
VeriFair [18]	Verification, Specification	how to verify fairness specifications, i.e., fairness properties of ML programs	post	adaptive concentration inequalities	Black
Justicia [65]	Verification	how to formally verify the fairness metrics are satisfied by different algorithms on different datasets	pre	stochastic satisfiability (SSAT)	Black
Checklist [123]	Testing	how to test (fairness) behaviors of NLP systems	post	behavioral testing, template -based test generation	Black
FairML [2]	Auditing, Analysis	how to determine the significance of inputs in assessing the fairness of black-box models	post	model compression, input ranking algorithms	Black
MT-NLP [106]	Testing, Mitigation	how to determine if NLP models are free of unfair bias toward certain sub-populations/groups	post	metamorphic testing	Black
ASTRAEA [137]	Debugging, Testing, Mitigation	how to perform fairness testing without an existing dataset, i.e., no training data access	post	grammar-based testing, metamorphic relations	Black
Aequitas [126]	Auditing, Analysis	how to audit for bias and fairness when developing and deploying algorithmic decision making systems	post	bias audit toolkit to support many bias metrics	Black
FairTest [146]	Testing, Debugging	how to detect unwarranted associations (UA) (disparate impact, offensive labels, and uneven error rates) between model outcomes and data attributes	post	unwarranted associations (UA) framework to determine UA between outcomes and attributes	Black
Themis-ml [16]	Fair Learning, Mitigation, Analysis, Auditing	how to measure, understand, and mitigate the implicit historical biases in socially sensitive data	pre, in & post	API for Fair ML for simple binary classifier	White, Black, & Grey
Fairify [26]	Verification Certification	how to verify individual fairness property in neural network (NN) models	in	SMT, formal analysis, pruning, input partitioning, interval arithmetic and activation heuristic	White

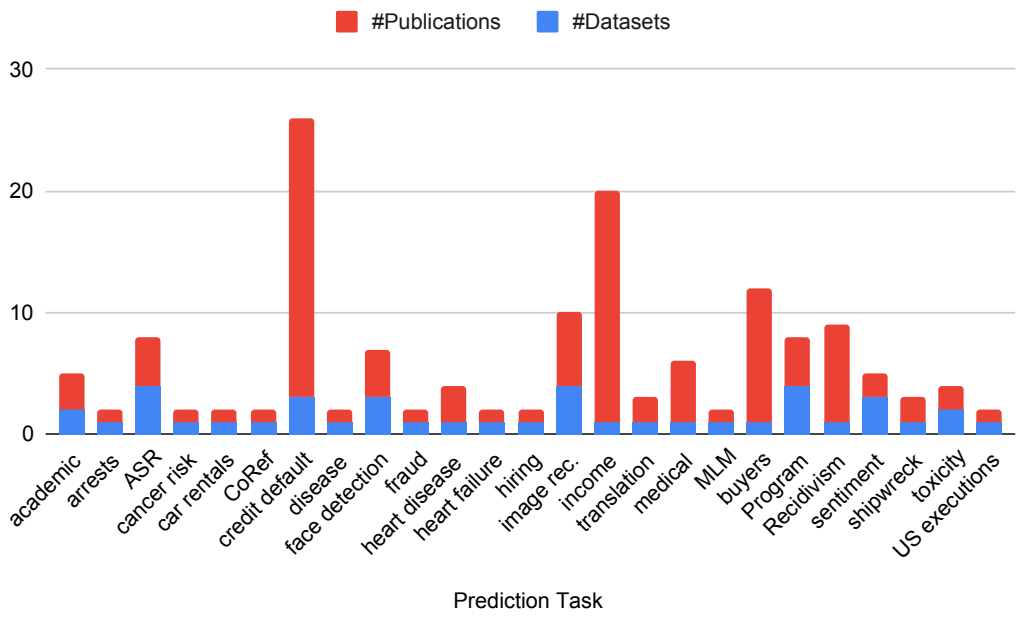


Fig. 14. Details of Publications and Datasets per Prediction Task

Table 6. Details of Open Problems and Future Research Opportunities (more recent (2022 and 2023) solutions are in bracket)

Open Problems	**Problem Description**	**Potential Solutions**	**Sample Related Work**
Fairness Test Metrics and Adequacy	Measuring when fairness testing is sufficient/enough	Design of fairness test metrics and adequacy criteria	[49, 53, 95, 105, 117, 135] [107] ([108, 168])
Automatic Repair of Biased Classifiers	How to automatically repair biased classifiers to be (less or) un-biased?	Automatic Program Repair for fairness property	[8, 128, 140] ([79, 142])
Tooling for Fairness Property Specification	Specifying and engineering fairness properties for learning-based systems	Requirement Engineering tool support for Fairness properties	[18, 133] ([17, 54])
Unexplored or Poorly Understood Biases	Analyzing rare biases (e.g., age), complex or intersectional biases (e.g., age × gender)?	Fairness Analysis Support for rare, complex or intersectional Biases	[30, 33] ([41, 68])
Sequential and Long-term Fairness concerns	How to analyse/maintain fairness as the AI system evolves over time?	Techniques to support analysis of sequential and long-term fairness	[165] ([67, 119])
Human factors in fairness analysis	E.g., evaluating the harm induced by fairness violations to humans/society	Empirical studies of Human Factors in Fairness Analysis	[44, 75, 98] ([55, 59, 60, 71, 114])
Non-Specific/Holistic mitigation approaches	Designing bias mitigation methods that are agnostic of tasks, domains or datasets	General (i.e., task, domain and dataset -agnostic) bias analysis techniques	[164]
Fair Policy, Legalisation, and Compliance	How to design fairness analysis tools for policy makers and compliance officers?	Fairness Analysis Tool Support for Policy and Compliance Analysis	[102, 115]

REFERENCES

[1] Serge Abiteboul and Julia Stoyanovich. 2019. Transparency, fairness, data protection, neutrality: Data management challenges in the face of new regulation. *Journal of Data and Information Quality (JDIQ)* 11, 3 (2019), 1–9.

[2] Julius A Adebayo et al. 2016. *FairML: ToolBox for diagnosing bias in predictive modeling*. Ph.D. Dissertation. Massachusetts Institute of Technology.

[3] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. 2018. A reductions approach to fair classification. In *International Conference on Machine Learning*. PMLR, 60–69.

[4] Chirag Agarwal, Himabindu Lakkaraju, and Marinka Zitnik. 2021. Towards a unified framework for fair and stable graph representation learning. In *Uncertainty in Artificial Intelligence*. PMLR, 2114–2124.

[5] Aniya Aggarwal, Pranay Lohia, Seema Nagar, Kuntal Dey, and Diptikalyan Saha. 2019. Black box fairness testing of machine learning models. In *Proceedings of the 2019 27th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. 625–635.

[6] Aniya Aggarwal, Samiulla Shaikh, Sandeep Hans, Swastik Haldar, Rema Ananthanarayanan, and Diptikalyan Saha. 2021. Testing framework for black-box AI models. In *2021 IEEE/ACM 43rd International Conference on Software Engineering: Companion Proceedings (ICSE-Companion)*. IEEE, 81–84.

[7] Aws Albarghouthi, Loris D'Antoni, Samuel Drews, and Aditya V Nori. 2017. Fairsquare: probabilistic verification of program fairness. *Proceedings of the ACM on Programming Languages* 1, OOPSLA (2017), 1–30.

[8] Aws Albarghouthi, Loris D'Antoni, and Samuel Drews. 2017. Repairing decision-making programs under uncertainty. In *International Conference on Computer Aided Verification*. Springer, 181–200.

[9] Aws Albarghouthi and Samuel Vinitsky. 2019. Fairness-aware programming. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 211–219.

[10] Jose Manuel Alvarez and Salvatore Ruggieri. 2023. Counterfactual Situation Testing: Uncovering Discrimination under Fairness given the Difference. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '23)*. Association for Computing Machinery, New York, NY, USA, Article 2, 11 pages. https://doi.org/10.1145/3617694.3623222

[11] Rico Angell, Brittany Johnson, Yuriy Brun, and Alexandra Meliou. 2018. Themis: Automatically testing software for discrimination. In *Proceedings of the 2018 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. 871–875.

[12] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2022. Machine bias. In *Ethics of data and analytics*. Auerbach Publications, 254–264.

[13] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Accessed:09.11.2023. Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[14] Muhammad Hilmi Asyrofi, Zhou Yang, Imam Nur Bani Yusuf, Hong Jin Kang, Ferdian Thung, and David Lo. 2021. Biasfinder: Metamorphic test generation to uncover bias for sentiment analysis systems. *IEEE Transactions on Software Engineering* (2021).

[15] Fatma Basak Aydemir and Fabiano Dalpiaz. 2018. A roadmap for ethics-aware software engineering. In *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*. IEEE, 15–21.

[16] Niels Bantilan. 2018. Themis-ml: A fairness-aware machine learning interface for end-to-end discrimination discovery and mitigation. *Journal of Technology in Human Services* 36, 1 (2018), 15–30.

[17] Luciano Baresi, Chiara Criscuolo, and Carlo Ghezzi. 2023. Understanding fairness requirements for ml-based software. In *2023 IEEE 31st International Requirements Engineering Conference (RE)*. IEEE, 341–346.

[18] Osbert Bastani, Xin Zhang, and Armando Solar-Lezama. 2019. Probabilistic verification of fairness properties via concentration. *Proceedings of the ACM on Programming Languages* 3, OOPSLA (2019), 1–27.

[19] Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilović, et al. 2019. AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development* 63, 4/5 (2019), 4–1.

[20] Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, et al. 2019. Think your artificial intelligence software is fair? Think again. *IEEE Software* 36, 4 (2019), 76–80.

[21] Nelly Bencomo, Jin LC Guo, Rachel Harrison, Hans-Martin Heyn, and Tim Menzies. 2021. The Secret to Better AI and Better Software (Is Requirements Engineering). *IEEE Software* 39, 1 (2021), 105–110.

[22] Richard Berk, Hoda Heidari, Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. 2017. A Convex Framework for Fair Regression. *Fairness, Accountability, and Transparency in Machine Learning* (2017).

[23] Sarah Bird, Miro Dudík, Richard Edgar, Brandon Horn, Roman Lutz, Vanessa Milan, Mehrnoosh Sameki, Hanna Wallach, and Kathleen Walker. 2020. Fairlearn: A toolkit for assessing and improving fairness in AI. *Microsoft, Tech. Rep. MSR-TR-2020-32* (2020).

[24] Sumon Biswas and Hridesh Rajan. 2020. Do the machine learning models on a crowd sourced platform exhibit bias? an empirical study on model fairness. In *Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. 642–653.

[25] Sumon Biswas and Hridesh Rajan. 2021. Fair Preprocessing: Towards Understanding Compositional Fairness of Data Transformers in Machine Learning Pipeline. *arXiv preprint arXiv:2106.06054* (2021).

[26] Sumon Biswas and Hridesh Rajan. 2023. Fairify: Fairness verification of neural networks. In *2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE)*. IEEE, 1546–1558.

[27] Emily Black, Samuel Yeom, and Matt Fredrikson. 2020. Fliptest: fairness testing via optimal transport. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 111–121.

[28] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 5454–5476.

[29] Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. 2021. Stereotyping Norwegian salmon: an inventory of pitfalls in fairness benchmark datasets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 1004–1015.

[30] Martim Brandao. 2019. Age and gender bias in pedestrian detection algorithms. In *Proceedings of the Workshop on Fairness Accountability Transparency and Ethics in Computer Vision at IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*.

[31] Yuriy Brun and Alexandra Meliou. 2018. Software fairness. In *Proceedings of the 2018 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. 754–759.

[32] Margaret Burnett, Simone Stumpf, Jamie Macbeth, Stephann Makri, Laura Beckwith, Irwin Kwan, Anicia Peters, and William Jernigan. 2016. GenderMag: A method for evaluating software's gender inclusiveness. *Interacting with Computers* 28, 6 (2016), 760–787.

[33] Ángel Alexander Cabrera, Will Epperson, Fred Hohman, Minsuk Kahng, Jamie Morgenstern, and Duen Horng Chau. 2019. FairVis: Visual analytics for discovering intersectional bias in machine learning. In *2019 IEEE Conference on Visual Analytics Science and Technology (VAST)*. IEEE, 46–56.

[34] Toon Calders and Sicco Verwer. 2010. Three naive Bayes approaches for discrimination-free classification. *Data mining and knowledge discovery* 21, 2 (2010), 277–292.

[35] Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. 2017. Optimized pre-processing for discrimination prevention. *Advances in neural information processing systems* 30 (2017).

[36] L Elisa Celis, Lingxiao Huang, Vijay Keswani, and Nisheeth K Vishnoi. 2019. Classification with fairness constraints: A meta-algorithm with provable guarantees. In *Proceedings of the conference on fairness, accountability, and transparency*. 319–328.

[37] Joymallya Chakraborty, Suvodeep Majumder, and Tim Menzies. 2021. Bias in machine learning software: why? how? what to do?. In *Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. 429–440.

[38] Joymallya Chakraborty, Suvodeep Majumder, Zhe Yu, and Tim Menzies. 2020. Fairway: A way to build fair ml software. In *Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. 654–665.

[39] Jialuo Chen, Jingyi Wang, Xingjun Ma, Youcheng Sun, Jun Sun, Peixin Zhang, and Peng Cheng. 2023. QuoTe: Quality-Oriented Testing for Deep Learning Systems. *ACM Trans. Softw. Eng. Methodol.* 32, 5, Article 125 (jul 2023), 33 pages. https://doi.org/10.1145/3582573

[40] Zhenpeng Chen, Jie M. Zhang, Federica Sarro, and Mark Harman. 2023. A Comprehensive Empirical Study of Bias Mitigation Methods for Machine Learning Classifiers. *ACM Trans. Softw. Eng. Methodol.* 32, 4, Article 106 (may 2023), 30 pages. https://doi.org/10.1145/3583561

[41] Zhenpeng Chen, Jie M Zhang, Federica Sarro, and Mark Harman. 2024. Fairness improvement with multiple protected attributes: How far are we? (2024), 1–13.

[42] Jürgen Cito, Isil Dillig, Seohyun Kim, Vijayaraghavan Murali, and Satish Chandra. 2021. Explaining mispredictions of machine learning models using rule induction. In *Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. 716–727.

[43] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*. 797–806.

[44] Jenna Cryan, Shiliang Tang, Xinyi Zhang, Miriam Metzger, Haitao Zheng, and Ben Y Zhao. 2020. Detecting gender stereotypes: lexicon vs. supervised learning methods. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–11.

[45] Anubrata Das and Matthew Lease. 2019. A Conceptual Framework for Evaluating Fairness in Search. *arXiv preprint arXiv:1907.09328* (2019).

[46] Terrance De Vries, Ishan Misra, Changhan Wang, and Laurens Van der Maaten. 2019. Does object recognition work for everyone?. In *Proceedings of the Workshop on Fairness Accountability Transparency and Ethics in Computer Vision at IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 52–59.

[47] Emily Denton, Ben Hutchinson, Margaret Mitchell, and Timnit Gebru. 2019. Detecting bias with generative counterfactual face attribute augmentation. (2019).

[48] Marina Drosou, HV Jagadish, Evaggelia Pitoura, and Julia Stoyanovich. 2017. Diversity in big data: A review. *Big data* 5, 2 (2017), 73–84.

[49] Xiaoning Du, Xiaofei Xie, Yi Li, Lei Ma, Yang Liu, and Jianjun Zhao. 2019. DeepStellar: Model-Based Quantitative Analysis of Stateful Deep Learning Systems. In *Proceedings of the 2019 27th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering* (Tallinn, Estonia) *(ESEC/FSE 2019)*. Association for Computing Machinery, New York, NY, USA, 477–487. https://doi.org/10.1145/3338906.3338954

[50] Ming Fan, Wenying Wei, Wuxia Jin, Zijiang Yang, and Ting Liu. 2022. Explanation-Guided Fairness Testing through Genetic Algorithm. In *2022 IEEE/ACM 44th International Conference on Software Engineering (ICSE)*. IEEE.

[51] Golnoosh Farnad, Behrouz Babaki, and Michel Gendreau. 2020. A unifying framework for fairness-aware influence maximization. In *Companion Proceedings of the Web Conference 2020*. 714–722.

[52] 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*. 259–268.

[53] Yang Feng, Qingkai Shi, Xinyu Gao, Jun Wan, Chunrong Fang, and Zhenyu Chen. 2020. DeepGini: Prioritizing Massive Tests to Enhance the Robustness of Deep Neural Networks. In *Proceedings of the 29th ACM SIGSOFT International Symposium on Software Testing and Analysis* (Virtual Event, USA) *(ISSTA 2020)*. Association for Computing Machinery, New York, NY, USA, 177–188. https://doi.org/10.1145/3395363.3397357

[54] Carmine Ferrara, Francesco Casillo, Carmine Gravino, Andrea De Lucia, and Fabio Palomba. 2024. Refair: Toward a context-aware recommender for fairness requirements engineering. (2024), 1–12.

[55] Carmine Ferrara, Giulia Sellitto, Filomena Ferrucci, Fabio Palomba, and Andrea De Lucia. 2023. Fairness-aware machine learning engineering: how far are we? *Empirical Software Engineering* 29, 1 (2023), 9.

[56] Anjalie Field, Su Lin Blodgett, Zeerak Waseem, and Yulia Tsvetkov. 2021. A Survey of Race, Racism, and Anti-Racism in NLP. *arXiv preprint arXiv:2106.11410* (2021).

[57] Anthony Finkelstein, Mark Harman, S Afshin Mansouri, Jian Ren, and Yuanyuan Zhang. 2009. A search based approach to fairness analysis in requirement assignments to aid negotiation, mediation and decision making. *Requirements engineering* 14, 4 (2009), 231–245.

[58] Jessie Finocchiaro, Roland Maio, Faidra Monachou, Gourab K Patro, Manish Raghavan, Ana-Andreea Stoica, and Stratis Tsirtsis. 2021. Bridging machine learning and mechanism design towards algorithmic fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 489–503.

[59] Claudia Flores-Saviaga, Christopher Curtis, and Saiph Savage. 2023. Inclusive Portraits: Race-Aware Human-in-the-Loop Technology. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '23)*. Association for Computing Machinery, New York, NY, USA, Article 15, 11 pages. https://doi.org/10.1145/3617694.3623235

[60] Aimen Gaba, Zhanna Kaufman, Jason Cheung, Marie Shvakel, Kyle Wm Hall, Yuriy Brun, and Cindy Xiong Bearfield. 2023. My Model is Unfair, Do People Even Care? Visual Design Affects Trust and Perceived Bias in Machine Learning. *IEEE Transactions on Visualization and Computer Graphics* (2023).

[61] Sainyam Galhotra, Yuriy Brun, and Alexandra Meliou. 2017. Fairness testing: testing software for discrimination. In *Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering*. 498–510.

[62] Xuanqi Gao, Juan Zhai, Shiqing Ma, Chao Shen, Yufei Chen, and Qian Wang. 2022. FairNeuron: Improving Deep Neural Network Fairness with Adversary Games on Selective Neurons. In *2022 IEEE/ACM 44th International Conference on Software Engineering (ICSE)*. IEEE.

[63] Xuanqi Gao, Juan Zhai, Shiqing Ma, Chao Shen, Yufei Chen, and Shiwei Wang. 2023. CILIATE: Towards Fairer Class-based Incremental Learning by Dataset and Training Refinement. *arXiv preprint arXiv:2304.04222* (2023).

[64] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Commun. ACM* 64, 12 (2021), 86–92.

[65] Bishwamittra Ghosh, Debabrota Basu, and Kuldeep S Meel. 2021. Justicia: A Stochastic SAT Approach to Formally Verify Fairness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 7554–7563.

[66] Stephen Giguere, Blossom Metevier, Yuriy Brun, Bruno Castro Da Silva, Philip S Thomas, and Scott Niekum. 2022. Fairness guarantees under demographic shift. In *Proceedings of the 10th International Conference on Learning Representations (ICLR)*.

[67] Usman Gohar, Sumon Biswas, and Hridesh Rajan. 2023. Towards understanding fairness and its composition in ensemble machine learning. In *2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE)*. IEEE, 1533–1545.

[68] Usman Gohar and Lu Cheng. 2023. A Survey on Intersectional Fairness in Machine Learning: Notions, Mitigation, and Challenges. *arXiv preprint arXiv:2305.06969* (2023).

[69] CV González Zelaya, P Missier, and D Prangle. 2019. Parametrised data sampling for fairness optimisation. In *2019 XAI Workshop at SIGKDD, Anchorage, AK, USA*.

[70] 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*. 903–912.

[71] Nina Grgić-Hlača, Gabriel Lima, Adrian Weller, and Elissa M. Redmiles. 2022. Dimensions of Diversity in Human Perceptions of Algorithmic Fairness. In *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '22)*. Association for Computing Machinery, New York, NY, USA, Article 21, 12 pages. https://doi.org/10.1145/3551624.3555306

[72] Emitzá Guzmán, Ricarda Anna-Lena Fischer, and Janey Kok. 2023. Mind the gap: gender, micro-inequities and barriers in software development. *Empirical Software Engineering* 29, 1 (2023), 17.

[73] Galen Harrison, Julia Hanson, Christine Jacinto, Julio Ramirez, and Blase Ur. 2020. An empirical study on the perceived fairness of realistic, imperfect machine learning models. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 392–402.

[74] Austin Hoag, James E. Kostas, Bruno Castro da Silva, Philip S. Thomas, and Yuriy Brun. 2023. Seldonian Toolkit: Building Software with Safe and Fair Machine Learning. In *2023 IEEE/ACM 45th International Conference on Software Engineering: Companion Proceedings (ICSE-Companion)*. 107–111. https://doi.org/10.1109/ICSE-Companion58688.2023.00035

[75] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudik, and Hanna Wallach. 2019. Improving fairness in machine learning systems: What do industry practitioners need?. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–16.

[76] Sara Hooker, Nyalleng Moorosi, Gregory Clark, Samy Bengio, and Emily Denton. 2020. Characterising bias in compressed models. *arXiv preprint arXiv:2010.03058* (2020).

[77] Max Hort, Rebecca Moussa, and Federica Sarro. 2023. Multi-objective search for gender-fair and semantically correct word embeddings. *Applied Soft Computing* 133 (2023), 109916.

[78] Max Hort, Jie M Zhang, Federica Sarro, and Mark Harman. 2021. Fairea: a model behaviour mutation approach to benchmarking bias mitigation methods. In *Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. 994–1006.

[79] Max Hort, Jie M. Zhang, Federica Sarro, and Mark Harman. 2024. Search-based Automatic Repair for Fairness and Accuracy in Decision-making Software. *Empirical Software Engineering* 29, 1 (2024), 36. https://doi.org/10.1007/s10664-023-10419-3

[80] Waqar Hussain, Davoud Mougouei, and Jon Whittle. 2018. Integrating social values into software design patterns. In *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*. IEEE, 8–14.

[81] Wiebke (Toussaint) Hutiri, Aaron Yi Ding, Fahim Kawsar, and Akhil Mathur. 2023. Tiny, Always-on, and Fragile: Bias Propagation through Design Choices in On-Device Machine Learning Workflows. *ACM Trans. Softw. Eng. Methodol.* 32, 6, Article 155 (sep 2023), 37 pages. https://doi.org/10.1145/3591867

[82] HV Jagadish, Julia Stoyanovich, and Bill Howe. 2021. Covid-19 brings data equity challenges to the fore. *Digital Government: Research and Practice* 2, 2 (2021), 1–7.

[83] Philips George John, Deepak Vijaykeerthy, and Diptikalyan Saha. 2020. Verifying individual fairness in machine learning models. In *Conference on Uncertainty in Artificial Intelligence*. PMLR, 749–758.

[84] Brittany Johnson, Jesse Bartola, Rico Angell, Sam Witty, Stephen Giguere, and Yuriy Brun. 2023. Fairkit, fairkit, on the wall, who's the fairest of them all? Supporting fairness-related decision-making. *EURO Journal on Decision Processes* 11 (2023), 100031. https://doi.org/10.1016/j.ejdp.2023.100031

[85] Brittany Johnson and Yuriy Brun. 2022. Fairkit-Learn: A Fairness Evaluation and Comparison Toolkit. In *Proceedings of the ACM/IEEE 44th International Conference on Software Engineering: Companion Proceedings* (Pittsburgh, Pennsylvania) *(ICSE '22)*. Association for Computing Machinery, New York, NY, USA, 70–74. https://doi.org/10.1145/3510454.3516830

[86] Faisal Kamiran and Toon Calders. 2009. Classifying without discriminating. In *2009 2nd international conference on computer, control and communication*. IEEE, 1–6.

[87] Faisal Kamiran and Toon Calders. 2012. Data preprocessing techniques for classification without discrimination. *Knowledge and information systems* 33, 1 (2012), 1–33.

[88] Faisal Kamiran, Toon Calders, and Mykola Pechenizkiy. 2010. Discrimination aware decision tree learning. In *2010 IEEE International Conference on Data Mining*. IEEE, 869–874.

[89] Faisal Kamiran, Asim Karim, and Xiangliang Zhang. 2012. Decision theory for discrimination-aware classification. In *2012 IEEE 12th International Conference on Data Mining*. IEEE, 924–929.

[90] Faisal Kamiran, Sameen Mansha, Asim Karim, and Xiangliang Zhang. 2018. Exploiting reject option in classification for social discrimination control. *Information Sciences* 425 (2018), 18–33.

[91] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. 2012. Fairness-aware classifier with prejudice remover regularizer. In *Joint European conference on machine learning and knowledge discovery in databases*. Springer, 35–50.

[92] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. 2018. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International Conference on Machine Learning*. PMLR, 2564–2572.

[93] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. 2019. An empirical study of rich subgroup fairness for machine learning. In *Proceedings of the conference on fairness, accountability, and transparency*. 100–109.

[94] Byungju Kim, Hyunwoo Kim, Kyungsu Kim, Sungjin Kim, and Junmo Kim. 2019. Learning not to learn: Training deep neural networks with biased data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9012–9020.

[95] Jinhan Kim, Robert Feldt, and Shin Yoo. 2019. Guiding deep learning system testing using surprise adequacy. In *2019 IEEE/ACM 41st International Conference on Software Engineering (ICSE)*. IEEE, 1039–1049.

[96] Michael P Kim, Amirata Ghorbani, and James Zou. 2019. Multiaccuracy: Black-box post-processing for fairness in classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 247–254.

[97] Caitlin Kuhlman, MaryAnn VanValkenburg, and Elke Rundensteiner. 2019. Fare: Diagnostics for fair ranking using pairwise error metrics. In *The World Wide Web Conference*. 2936–2942.

[98] Michelle Seng Ah Lee and Jat Singh. 2021. The landscape and gaps in open source fairness toolkits. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–13.

[99] Chenglu Li, Wanli Xing, and Walter Leite. 2021. Yet Another Predictive Model? Fair Predictions of Students' Learning Outcomes in an Online Math Learning Platform. In *LAK21: 11th International Learning Analytics and Knowledge Conference*. 572–578.

[100] Xinyue Li, Zhenpeng Chen, Jie M Zhang, Federica Sarro, Ying Zhang, and Xuanzhe Liu. 2023. Dark-skin individuals are at more risk on the street: Unmasking fairness issues of autonomous driving systems. *arXiv preprint arXiv:2308.02935* (2023).

[101] Yanhui Li, Linghan Meng, Lin Chen, Li Yu, Di Wu, Yuming Zhou, and Baowen Xu. 2022. Training Data Debugging for the Fairness of Machine Learning Software. In *2022 IEEE/ACM 44th International Conference on Software Engineering (ICSE)*. IEEE.

[102] Lizhen Liang and Daniel E Acuna. 2020. Artificial mental phenomena: Psychophysics as a framework to detect perception biases in AI models. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 403–412.

[103] Zachary Lipton, Julian McAuley, and Alexandra Chouldechova. 2018. Does mitigating ML's impact disparity require treatment disparity? *Advances in neural information processing systems* 31 (2018).

[104] Kristian Lum and Tarak Shah. 2019. Measures of fairness for New York City's Supervised Release Risk Assessment Tool. *Human Rights Data Analytics Group* (2019), 21.

[105] Lei Ma, Fuyuan Zhang, Jiyuan Sun, Minhui Xue, Bo Li, Felix Juefei-Xu, Chao Xie, Li Li, Yang Liu, Jianjun Zhao, et al. 2018. Deepmutation: Mutation testing of deep learning systems. In *2018 IEEE 29th International Symposium on Software Reliability Engineering (ISSRE)*. IEEE, 100–111.

[106] Pingchuan Ma, Shuai Wang, and Jin Liu. 2020. Metamorphic Testing and Certified Mitigation of Fairness Violations in NLP Models.. In *IJCAI*. 458–465.

[107] Wei Ma, Mike Papadakis, Anestis Tsakmalis, Maxime Cordy, and Yves Le Traon. 2021. Test Selection for Deep Learning Systems. *ACM Trans. Softw. Eng. Methodol.* 30, 2, Article 13 (Jan. 2021), 22 pages. https://doi.org/10.1145/3417330

[108] Suvodeep Majumder, Joymallya Chakraborty, Gina R. Bai, Kathryn T. Stolee, and Tim Menzies. 2023. Fair Enough: Searching for Sufficient Measures of Fairness. *ACM Trans. Softw. Eng. Methodol.* 32, 6, Article 134 (sep 2023), 22 pages. https://doi.org/10.1145/3585006

[109] Blossom Metevier, Stephen Giguere, Sarah Brockman, Ari Kobren, Yuriy Brun, Emma Brunskill, and Philip Thomas. 2019. Offline contextual bandits with high probability fairness guarantees. *Advances in neural information processing systems* 32 (2019).

[110] Verya Monjezi, Ashutosh Trivedi, Gang Tan, and Saeid Tizpaz-Niari. 2023. Information-Theoretic Testing and Debugging of Fairness Defects in Deep Neural Networks. *arXiv preprint arXiv:2304.04199* (2023).

[111] Daniel Perez Morales, Takashi Kitamura, and Shingo Takada. 2021. Coverage-Guided Fairness Testing. In *International Conference on Intelligence Science*. Springer, 183–199.

[112] Aniruddhan Murali, Gaurav Sahu, Kishanthan Thangarajah, Brian Zimmerman, Gema Rodríguez-Pérez, and Meiyappan Nagappan. 2024. Diversity in issue assignment: humans vs bots. *Empirical Software Engineering* 29, 2 (2024), 37.

[113] Giang Nguyen, Sumon Biswas, and Hridesh Rajan. 2023. Fix Fairness, Don't Ruin Accuracy: Performance Aware Fairness Repair using AutoML. *arXiv preprint arXiv:2306.09297* (2023).

[114] Julian Nyarko, Sharad Goel, and Roseanna Sommers. 2021. Breaking Taboos in Fair Machine Learning: An Experimental Study. In *Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization* (–, NY, USA) *(EAAMO '21)*. Association for Computing Machinery, New York, NY, USA, Article 14, 11 pages. https://doi.org/10.1145/3465416.3483291

[115] U.S. Department of Labor. Accessed: 25.04.2022. Affirmative Action. https://www.dol.gov/general/topic/hiring/affirmativeact

[116] Moses Openja, Gabriel Laberge, and Foutse Khomh. 2023. Detection and evaluation of bias-inducing features in machine learning. *Empirical Software Engineering* 29, 1 (2023), 22.

[117] Kexin Pei, Yinzhi Cao, Junfeng Yang, and Suman Jana. 2017. Deepxplore: Automated whitebox testing of deep learning systems. In *proceedings of the 26th Symposium on Operating Systems Principles*. 1–18.

[118] Kewen Peng, Joymallya Chakraborty, and Tim Menzies. 2022. FairMask: Better fairness via model-based rebalancing of protected attributes. *IEEE Transactions on Software Engineering* 49, 4 (2022), 2426–2439.

[119] Anjana Perera, Aldeida Aleti, Chakkrit Tantithamthavorn, Jirayus Jiarpakdee, Burak Turhan, Lisa Kuhn, and Katie Walker. 2022. Search-based fairness testing for regression-based machine learning systems. *Empirical Software Engineering* 27, 3 (2022), 79.

[120] Nga Pham, Hung Pham-Ngoc, and Anh Nguyen-Duc. 2023. Fairness Requirement in AI Engineering–A Review on Current Research and Future Directions. In *International Conference on Sustainability in Software Engineering & Business Information Management*. Springer, 3–13.

[121] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. 2017. On fairness and calibration. *Advances in neural information processing systems* 30 (2017).

[122] Sai Sathiesh Rajan, Sakshi Udeshi, and Sudipta Chattopadhyay. 2022. AequeVox: Automated Fairness Testing of Speech Recognition Systems. (2022).

[123] Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond Accuracy: Behavioral Testing of NLP Models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 4902–4912.

[124] Luke Rodriguez, Babak Salimi, Haoyue Ping, Julia Stoyanovich, and Bill Howe. 2018. MobilityMirror: Bias-adjusted transportation datasets. In *Workshop on Big Social Data and Urban Computing*. Springer, 18–39.

[125] Debjani Saha, Candice Schumann, Duncan C McElfresh, John P Dickerson, Michelle L Mazurek, and Michael Carl Tschantz. 2020. Human comprehension of fairness in machine learning. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 152–152.

[126] Pedro Saleiro, Benedict Kuester, Loren Hinkson, Jesse London, Abby Stevens, Ari Anisfeld, Kit T Rodolfa, and Rayid Ghani. 2018. Aequitas: A bias and fairness audit toolkit. *arXiv preprint arXiv:1811.05577* (2018).

[127] Babak Salimi, Bill Howe, and Dan Suciu. 2019. Data management for causal algorithmic fairness. *arXiv preprint arXiv:1908.07924* (2019).

[128] Babak Salimi, Bill Howe, and Dan Suciu. 2020. Database repair meets algorithmic fairness. *ACM SIGMOD Record* 49, 1 (2020), 34–41.

[129] Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A Smith, and Yejin Choi. 2020. Social Bias Frames: Reasoning about Social and Power Implications of Language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 5477–5490.

[130] Prasanna Sattigeri, Samuel C Hoffman, Vijil Chenthamarakshan, and Kush R Varshney. 2019. Fairness GAN: Generating datasets with fairness properties using a generative adversarial network. *IBM Journal of Research and Development* 63, 4/5 (2019), 3–1.

[131] Daniel Schwarcz. 2021. Health-Based Proxy Discrimination, Artificial Intelligence, and Big Data. *Artificial Intelligence, and Big Data (March 3, 2021). Houston Journal of Health Law and Policy* (2021).

[132] Shahar Segal, Yossi Adi, Benny Pinkas, Carsten Baum, Chaya Ganesh, and Joseph Keshet. 2021. Fairness in the eyes of the data: Certifying machine-learning models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 926–935.

[133] Arnab Sharma, Caglar Demir, Axel-Cyrille Ngonga Ngomo, and Heike Wehrheim. 2021. MLCheck-Property-Driven Testing of Machine Learning Models. *arXiv preprint arXiv:2105.00741* (2021).

[134] Shubham Sharma, Yunfeng Zhang, Jesús M Ríos Aliaga, Djallel Bouneffouf, Vinod Muthusamy, and Kush R Varshney. 2020. Data augmentation for discrimination prevention and bias disambiguation. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 358–364.

[135] Weijun Shen, Yanhui Li, Lin Chen, Yuanlei Han, Yuming Zhou, and Baowen Xu. 2020. Multiple-Boundary Clustering and Prioritization to Promote Neural Network Retraining. In *2020 35th IEEE/ACM International Conference on Automated Software Engineering (ASE)*. 410–422.

[136] Kacper Sokol, Alexander Hepburn, Rafael Poyiadzi, Matthew Clifford, Raul Santos-Rodriguez, and Peter Flach. 2020. FAT forensics: a python toolbox for implementing and deploying fairness, accountability and transparency algorithms in predictive systems. *Journal of Open Source Software* 5, 49 (2020), 1904.

[137] Ezekiel Soremekun, Sakshi Udeshi, and Sudipta Chattopadhyay. 2022. Astraea: Grammar-based fairness testing. *IEEE Transactions on Software Engineering* (2022).

[138] Julia Stoyanovich. 2019. TransFAT: Translating fairness, accountability and transparency into data science practice. In *CEUR Workshop Proceedings*, Vol. 2417. CEUR-WS.

[139] Julia Stoyanovich, Serge Abiteboul, and Gerome Miklau. 2016. Data, responsibly: Fairness, neutrality and transparency in data analysis. In *International Conference on Extending Database Technology*.

[140] Zeyu Sun, Jie M Zhang, Mark Harman, Mike Papadakis, and Lu Zhang. 2020. Automatic testing and improvement of machine translation. In *Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering*. 974–985.

[141] Zeyu Sun, Jie M Zhang, Yingfei Xiong, Mark Harman, Mike Papadakis, and Lu Zhang. 2022. Improving Machine Translation Systems via Isotopic Replacement. In *2022 IEEE/ACM 44th International Conference on Software Engineering (ICSE)*. IEEE.

[142] Guanhong Tao, Weisong Sun, Tingxu Han, Chunrong Fang, and Xiangyu Zhang. 2022. RULER: discriminative and iterative adversarial training for deep neural network fairness. In *Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. 1173–1184.

[143] Philip S. Thomas, Bruno Castro da Silva, Andrew G. Barto, Stephen Giguere, Yuriy Brun, and Emma Brunskill. 2019. Preventing undesirable behavior of intelligent machines. *Science* 366, 6468 (2019), 999–1004. https://doi.org/10.1126/science.aag3311 arXiv:https://www.science.org/doi/pdf/10.1126/science.aag3311

[144] Yuchi Tian, Ziyuan Zhong, Vicente Ordonez, Gail Kaiser, and Baishakhi Ray. 2020. Testing DNN image classifiers for confusion & bias errors. In *Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering*. 1122–1134.

[145] Saeid Tizpaz-Niari, Ashish Kumar, Gang Tan, and Ashutosh Trivedi. 2022. Fairness-aware Configuration of Machine Learning Libraries. In *2022 IEEE/ACM 44th International Conference on Software Engineering (ICSE)*. IEEE.

[146] Florian Tramer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, Jean-Pierre Hubaux, Mathias Humbert, Ari Juels, and Huang Lin. 2017. Fairtest: Discovering unwarranted associations in data-driven applications. In *2017 IEEE European Symposium on Security and Privacy (EuroS&P)*. IEEE, 401–416.

[147] Sakshi Udeshi, Pryanshu Arora, and Sudipta Chattopadhyay. 2018. Automated directed fairness testing. In *Proceedings of the 33rd ACM/IEEE International Conference on Automated Software Engineering*. 98–108.

[148] Sahil Verma and Julia Rubin. 2018. Fairness definitions explained. In *2018 ieee/acm international workshop on software fairness (fairware)*. IEEE, 1–7.

[149] Yuxuan Wan, Wenxuan Wang, Pinjia He, Jiazhen Gu, Haonan Bai, and Michael R Lyu. 2023. Biasasker: Measuring the bias in conversational ai system. In *Proceedings of the 31st ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. 515–527.

[150] Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, and Vicente Ordonez. 2019. Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 5310–5319.

[151] Zeyu Wang, Klint Qinami, Ioannis Christos Karakozis, Kyle Genova, Prem Nair, Kenji Hata, and Olga Russakovsky. 2020. Towards fairness in visual recognition: Effective strategies for bias mitigation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 8919–8928.

[152] Nimmi Rashinika Weeraddana, Xiaoyan Xu, Mahmoud Alfadel, Shane McIntosh, and Meiyappan Nagappan. 2023. An empirical comparison of ethnic and gender diversity of DevOps and non-DevOps contributions to open-source projects. *Empirical Software Engineering* 28, 6 (2023), 150.

[153] Benjamin Wilson, Judy Hoffman, and Jamie Morgenstern. 2019. Predictive inequity in object detection. (2019).

[154] Yongkai Wu, Lu Zhang, Xintao Wu, and Hanghang Tong. 2019. Pc-fairness: A unified framework for measuring causality-based fairness. *Advances in Neural Information Processing Systems* 32 (2019).

[155] Yisong Xiao, Aishan Liu, Tianlin Li, and Xianglong Liu. 2023. Latent Imitator: Generating Natural Individual Discriminatory Instances for Black-Box Fairness Testing. *arXiv preprint arXiv:2305.11602* (2023).

[156] Zhou Yang, Muhammad Hilmi Asyrofi, and David Lo. 2021. BiasRV: Uncovering biased sentiment predictions at runtime. In *Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. 1540–1544.

[157] Jun Yu, Xinlong Hao, Haonian Xie, and Ye Yu. 2020. Fair face recognition using data balancing, enhancement and fusion. In *European Conference on Computer Vision*. Springer, 492–505.

[158] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. 2013. Learning fair representations. In *International conference on machine learning*. PMLR, 325–333.

[159] Jie M Zhang and Mark Harman. 2021. "Ignorance and Prejudice" in Software Fairness. In *2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE)*. IEEE, 1436–1447.

[160] Lingfeng Zhang, Yueling Zhang, and Min Zhang. 2021. Efficient white-box fairness testing through gradient search. In *Proceedings of the 30th ACM SIGSOFT International Symposium on Software Testing and Analysis*. 103–114.

[161] Mengdi Zhang and Jun Sun. 2022. Adaptive fairness improvement based on causality analysis. In *Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. 6–17.

[162] Mengdi Zhang, Jun Sun, Jingyi Wang, and Bing Sun. 2023. TestSGD: Interpretable Testing of Neural Networks against Subtle Group Discrimination. *ACM Trans. Softw. Eng. Methodol.* 32, 6, Article 137 (sep 2023), 24 pages. https://doi.org/10.1145/3591869

[163] Peixin Zhang, Jingyi Wang, Jun Sun, Guoliang Dong, Xinyu Wang, Xingen Wang, Jin Song Dong, and Ting Dai. 2020. White-box fairness testing through adversarial sampling. In *Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering*. 949–960.

[164] Peixin Zhang, Jingyi Wang, Jun Sun, Xinyu Wang, Guoliang Dong, Xingen Wang, Ting Dai, and Jin Song Dong. 2021. Automatic Fairness Testing of Neural Classifiers through Adversarial Sampling. *IEEE Transactions on Software Engineering* (2021).

[165] Xueru Zhang and Mingyan Liu. 2021. Fairness in learning-based sequential decision algorithms: A survey. In *Handbook of Reinforcement Learning and Control*. Springer, 525–555.

[166] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv preprint arXiv:1707.09457* (2017).

[167] Haibin Zheng, Zhiqing Chen, Tianyu Du, Xuhong Zhang, Yao Cheng, Shouling Ji, Jingyi Wang, Yue Yu, and Jinyin Chen. 2022. NeuronFair: Interpretable White-Box Fairness Testing through Biased Neuron Identification. In *2022 IEEE/ACM 44th International Conference on Software Engineering (ICSE)*. IEEE.

[168] W. Zheng, L. Lin, X. Wu, and X. Chen. 5555. An Empirical Study on Correlations between Deep Neural Network Fairness and Neuron Coverage Criteria. *IEEE Transactions on Software Engineering* 01 (jan 5555), 1–22. https://doi.org/10.1109/TSE.2023.3349001

[169] Indre Žliobaite, Faisal Kamiran, and Toon Calders. 2011. Handling conditional discrimination. In *2011 IEEE 11th International Conference on Data Mining*. IEEE, 992–1001.