附录 A(一级研究文献集合)：

[A1] Muller ST, Hoffman RR, Clancey W, Emrey A, Klein G. Explanation in human-AI systems: A literature meta-review, synopsis of key ideas and publications, and bibliography for explainable AI. arXiv preprint arXiv:1902.01876. 2019.

[A2] Mohseni S, Zarei N, Ragan ED. A multidisciplinary survey and framework for design and evaluation of explainable AI systems. ACM Transactions on Interactive Intelligent Systems, 2021,11(3-4):1-45.

[A3] Hoffman RR, Mueller ST, Klein G, Litman J. Metrics for explainable AI: challenges and prospects. arXiv preprint arXiv: 1812.04608. 2018.

[A4] Kaur D, Uslu S, Durresi A, Badve S, Dundar M. Trustworthy explainability acceptance: A new metric to measure the trustworthiness of interpretable AI medical diagnostic systems. In: Conference on Complex, Intelligent, and Software Intensive Systems. Spring, 2021. 35-46.

[A5] Ribera M, Lapedriza A. Can we do better explanations? A proposal of user-centered explainable AI. In: IUI workshops, 2019, 2327: 38.

[A6] Uslu S, Kaur D, Rivera SJ, Durresi A, Durresi M, Babbar-Sebens M. Trustworthy acceptance: a new metric for trustworthy artificial intelligence used in decision making in food-energy-water sectors. In: International Conference on Advanced Information Networking and Applications. Springer, Cham, 2021. 208-219.

[A7] Doshi-Velez F, Kim B. Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv: 1702.08608. 2017.

[A8] Arya V, Bellamy RK, Chen PY, Dhurandhar A, Hind M, Hoffman SC, Houde S, Liao QV, Luss R, Mojsilović A, Mourad S, Pedemonte P, Raghavendra R, Richards J, Sattigeri P, Shanmugam K, Singh M, Varshney KR, Wei D, Zhang Y. One explanation does not fit all: A toolkit and taxonomy of AI explainability techniques. arXiv preprint arXiv: 1909.03012. 2019.

[A9] Sokol K, Flach P. Explainability fact sheets: a framework for systematic assessment of explainable approaches. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. ACM, 2020. 56-67.

[A10] Dunn C, Moustafa N, Turnbull B. Robustness evaluations of sustainable machine learning models against learning models again data poisoning attacks in the internet of things. Sustainability, 2020,12(16):6434.

[A11] Sadeghzadeh AM, Shiravi S, Jalili R. Adversarial network traffic: Towards evaluating the robustness of deep-learning-based network traffic classification. IEEE Transactions on Network and Service Management, 2021,18(2):1962-1976.

[A12] Biggio B, Fumera G, Roli F. Security evaluation of pattern classifiers under attack. IEEE Transactions on Knowledge and Data Engineering, 2013,26(4):984-996.

[A13] Katzir Z, Elovici Y. Quantifying the resilience of machine learning classifiers used for cyber security. Expert System with Applications, 2018,92:419-429.

[A14] Guo J, Jiang Y, Zhao Y, Chen Q, Sun J. Dlfuzz: Differential fuzzing testing of deep learning systems. In: Proceedings of the 2018 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering. ACM, 2018. 739-743.

[A15] Carlini N, Athalye A, Papernot N, Brendel W, Rauber J, Tsipras D, Goodfellow I, Madry A, Kurakin A. On evaluating adversarial robustness. arXiv preprint arXiv: 1902.06705. 2019.

[A16] Croce F, Hein M. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In: International Conference on Machine Learning. PMLR, 2020. 2206-2216.

[A17] Goodfellow I. New CleverHans feature: Better adversarial robustness evaluations with attack bundling. arXiv preprint arXiv: 1811.03685. 2018.

[A18] Reagen B, Gupta U, Pentecost L, Whatmough P, Lee SK, Mulholland N, Brooks D, Wei GY. Ares: A framework for quantifying the resilience of deep neural networks. In: 2018 55th ACM/ESDA/IEEE Design Automation Conference (DAC). IEEE, 2018. 1-6.

[A19] Mahmoud A, Aggarwal N, Nobbe A, Vicarte JRS, Adve SV, Fletcher CW, Frosio I, Hari SKS. Pytorchfi: A runtime perturbation tool for dnns. In: 2020 50th Annual IEEE/IFIP International Conference on Dependable Systems and Networks Workshops (DSN-W). IEEE, 2020. 25-31.

[A20] Sun Y, Huang X, Kroening D, Sharp J, Hill M, Ashmore R. DeepConcolic: Testing and debugging deep neural networks. In: 2019

IEEE/ACM 41st International Conference on Software Engineering: Companion Proceedings (ICSE-Companion). IEEE, 2019. 111-114.

[A21] Zhang F, Chowdhury SP, Christakis M. Deepsearch: A simple and effective blackbox fuzzing of deep neural networks. In: Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering. ACM, 2020. 800-812.

[A22] Raj S, Jha SK, Ramanathan A, Pullum LL. Work-in-progress: testing autonomous cyber-physical systems using fuzzing features from convolutional neural networks. In: 2017 International Conference on Embedded Software (EMSOFT). IEEE, 2017. 1-2.

[A23] Sun Y, Wu M, Ruan W, Huang X, Kwiatkowska M, Kroening D. Concolic testing for deep neural networks. In: Proceedings of the 33rd ACM/IEEE International Conference on Automated Software Engineering. ACM, 2018. 109-119.

[A24] Chen Z, Narayanan N, Fang B, Li G, Pattabiraman K, DeBardeleben N. Tensorfi: A flexible fault injection framework for tensorflow applications. In: 2020 IEEE 31st International Symposium on Software Reliability Engineering (ISSRE). IEEE, 2020. 426-435.

[A25] Odena A, Olsson C, Andersen D, Goodfellow I. Tensorfuzz: Debugging neural networks with coverage-guided fuzzing. In: International Conference on Machine Learning. PMLR, 2019. 4901-4911.

[A26] Feldman M, Friedler SA, Moeller J, Scheidegger C, Venkatasubramanian S. Certifying and removing disparate impact. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2015. 259-268.

[A27] Calders T, Verwer S. Three naive Bayes approaches for discrimination-free classification. Data mining and knowledge discovery. 2010,21(2):277-292.

[A28] Hardt M, Price E, Srebro N. Equality of opportunity in supervised learning. Advances in neural information processing systems. 2016,29:3315-3323.

[A29] Dwork C, Hardt M, Pitassi T, Reingold O, Zemel R. Fairness through awareness. In: Proceedings of the 3rd innovations in theoretical computer science conference. ACM,2012. 214-226.

[A30] Chouldechova A, Roth A. The frontiers of fairness in machine learning. arXiv preprint. arXiv:1810.08810. 2018.

[A31] McSherry F. Privacy integrated queries: An extensible platform for privacy-preserving data analysis. In: Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data. ACM, 2009. 19-30.

[A32] Abadi M, Chu A, Goodfellow I, McMahan HB, Mironov I, Talwar K, Zhang L. Deep learning with differential privacy. In: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security. ACM, 2016. 308-318.

[A33] Long Y, Bindschaedler V, Gunter CA. Towards measuring membership privacy. arXiv preprint arXiv:1712.09136. 2017.

[A34] Huang L, Joseph AD, Nelson B, Rubinstein BI, Tygar JD. Adversarial machine learning. In: Proceedings 4th ACM Workshop Security and Artificial Intelligence. ACM, 2011. 43-58.

[A35] Rubinstein BI, Nelson B, Huang L, Joseph AD, Lau S H, Rao S, Tygar JD. Antidote: understanding and defending against poisoning of anomaly detectors. In: Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement. ACM, 2009. 1-14.

[A36] Demontis A, Melis M, Biggio B, Maiorca D, Arp D, Rieck K, Corona I, Giacinto G, Roil F. Yes, machine learning can be more secure! a case study on android malware detection. IEEE Transactions on Dependable and Secure Computing. 2017,16(4):711-724.

[A37] Nelson B, Barreno M, Jack Chi F, Joseph AD, Rubinstein BI, Saini U, Sutton C, Tygar JD, Xia K. Misleading learners: Co-opting your spam filter. In: Machine learning in cyber trust. Springer, 2009. 17-51.

[A38] Shen S, Tople S, Saxena P. AUror: Defending against poisoning attacks in collaborative deep learning systems. In: Proceedings of the 32nd Annual Conference on Computer Security Applications. ACM, 2016. 508–519.

[A39] Steinhardt J, Koh PWW, Liang PS. Certified defenses for data poisoning attacks. Advances in Neural Information Processing Systems. 2017: 3517–3529.

[A40] Baracaldo N, Chen B, Ludwig H, Safavi JA. Mitigating poisoning attacks on machine learning models: A data provenance-based approach. In: Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security. ACM, 2017. 103–110.

[A41] Aman MN, Chua KC, Sikdar B. Secure data provenance for the internet of things. In: Proceedings of the 3rd ACM Int'l Workshop on IoT Privacy, Trust, and Security. 2017: 11-14.

[A42] Zhang X, Zhu X, Wright S. Training set debugging using trusted items. In: Proceedings of the 32nd AAAI Conference on Artificial

Intelligence, 2018. 32(1): 4482-4489.

[A43] Koh PW, Liang P. Understanding black-box predictions via influence functions. In: International conference on machine learning. PMLR, 2017: 1885-1894.

[A44] Cretu GF, Stavrou A, Locasto ME, Stolfo SJ, Keromytis AD. Casting out demons: Sanitizing training data for anomaly sensors. In: 2008 IEEE Symposium on Security and Privacy. IEEE, 2008. 81-95.

[A45] Nelson B, Barreno M, Chi FJ, Joseph AD, Rubinstein BI, Saini U, Tygar JD, Xia K. Exploiting machine learning to subvert your spam filter. LEET, 2008,8(1):9.

[A46] Chen B, Carvalho W, Baracaldo N, Ludwig H, Edwards B, Lee T, Molloy I, Srivastava B. Detecting backdoor attacks on deep neural networks by activation clustering. arXiv preprint arXiv:1811.03728. 2018.

[A47] Wang B, Yao Y, Shan S, Li H, Viswanath B, Zheng H, Zhao BY. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks. In: 2019 IEEE Symposium on Security and Privacy (SP). IEEE, 2019. 707-723.

[A48] Carlini N, Wagner D. Towards evaluating the robustness of neural networks. In: Proceedings of the 2017 IEEE Symposium on Security and Privacy (SP). IEEE, 2017. 39-57.

[A49] Carlini N, Wagner D. Adversarial examples are not easily detected: Bypassing ten detection methods. In: Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security. ACM, 2017. 3-14

[A50] Meng D, Chen H. MagNet: a two-pronged defense against adversarial examples. In: Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security. ACM, 2017. 135-147.

[A51] Metzen JH, Genewein T, Fischer V, Bischoff B. On detecting adversarial perturbations. arXiv preprint arXiv:1702.04267. 2017.

[A52] Lu J, Issaranon T, Forsyth D. Safetynet: detecting and rejecting adversarial examples robustly. In: Proceedings of the IEEE International Conference on Computer Vision. IEEE, 2018. 446-454.

[A53] Hendrycks D, Gimpel K. Early methods for detecting adversarial images. arXiv preprint arXiv:1608.00530. 2016.

[A54] Song Y, Kim T, Nowozin S, Ermon S, Kushman N. Pixeldefend: Leveraging generative models to understand and defend against adversarial examples. arXiv preprint arXiv:1710.10766. 2017.

[A55] Biggio B, Roli F. Wild patterns: Ten years after the rise of adversarial machine learning. Pattern Recognition. 2018,84:317–31.

[A56] Dasgupta P, Collins J. A survey of game theoretic approaches for adversarial machine learning in cybersecurity tasks. AI Magazine. 2019,40(2):31-43.

[A57] Wang X, Li J, Kuang X, Tan YA, Li J. The security of machine learning in an adversarial setting: A survey. Journal of Parallel and Distributed Computing. 2019,130:12-23.

[A58] Wagner D, Soto P. Mimicry attacks on host-based intrusion detection systems. In: Proceedings of the ACM Conference on Computer and Communications Security. ACM, 2002. 255-64.

[A59] Shaham U, Yamada Y, Negahban S. Understanding adversarial training: Increasing local stability of supervised models through robust optimization. Neurocomputing. 2018,307:195-204.

[A60] Goodfellow IJ, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572. 2014.

[A61] Lee H, Han S, Lee J. Generative adversarial trainer: Defense to adversarial perturbations with gan. arXiv preprint arXiv:1705.03387. 2017.

[A62] Kurakin A, Goodfellow I, Bengio S. Adversarial machine learning at scale. arXiv preprint arXiv:1611.01236. 2016.

[A63] Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I, Fergus R. Intriguing properties of neural networks. arXiv preprint arXiv: 1312.6199. 2013.

[A64] Moosavi-Dezfooli SM, Fawzi A, Frossard P. Deepfool: A simple and accurate method to fool deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2016. 2574–2582.

[A65] Li Y, Li L, Wang L, Zhang T, Gong B. NATTACK: Learning the distributions of adversarial examples for an improved black-box attack on deep neural networks. In: International Conference on Machine Learning. PMLR, 2019. 3866-3876.

[A66] Tramèr F, Kurakin A, Papernot N, Goodfellow I, Boneh D, McDaniel P. Ensemble adversarial training: Attacks and defenses. arXiv preprint arXiv: 1705.07204. 2017.

[A67] Kannan H, Kurakin A, Goodfellow I. Adversarial logit pairing. arXiv preprint arXiv: 1803.06373. 2018.

[A68] Huang R, Xu B, Schuurmans D, Szepesvári C. Learning with a strong adversary. arXiv preprint arXiv:1511.03034. 2015.

[A69] Moosavi-Dezfooli SM, Fawzi A, Fawzi O, Frossard P. Universal adversarial perturbations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017. 1765-1773.

[A70] Papernot N, McDaniel P, Wu X, Jha S, Swami A. Distillation as a defense to adversarial perturbations against deep neural networks. In: 2016 IEEE Symposium on Security and Privacy. IEEE, 2016. 582-597.

[A71] Sethi TS, Kantardzic K, Ryu JW. 'Security theater': on the vulnerability of classifiers to exploratory attacks. In: Proceedings of the Pacific-Asia Workshop on Intelligence and Security Informatics. Springer, 2017. 49-63.

[A72] Papernot N, McDaniel P. On the effectiveness of defensive distillation. arXiv preprint arXiv:1607.05113. 2016.

[A73] Liang B, Li H, Su M, Li X, Shi W, Wang X. Detecting adversarial image examples in deep networks with adaptive noise reduction. IEEE Transactions on Dependable and Secure Computing. 2018,18(1):72-85.

[A74] Zantedeschi V, Nicolae MI, Rawat A. Efficient defenses against adversarial attacks. In: Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security. ACM, 2017. 39-49.

[A75] Xie C, Wang J, Zhang Z, Ren Z, Yuille A. Mitigating adversarial effects through randomization. arXiv preprint arXiv:1711.01991. 2017.

[A76] Lyu C, Huang K, Liang HN. A unified gradient regularization family for adversarial examples. In: 2015 IEEE International Conference on Data Mining. IEEE, 2016. 301-309.

[A77] Nguyen L, Wang S, Sinha A. A learning and masking approach to secure learning. In: International Conference on Decision and Game Theory for Security. Springer, 2018. 453-464.

[A78] Madry A, Makelov A, Schmidt L, Tsipras D, Vladu A. Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083. 2017.

[A79] Ross A, Doshi-Velez F. Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients. In: Proceedings of the 32nd AAAI Conference on Artificial Intelligence. AAAI, 2018. 1660-1669.

[A80] Moosavi-Dezfooli SM, Fawzi A, Uesato J, Frossard P. Robustness via curvature regularization, and vice versa. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, 2019. 9078-9086.

[A81] Gu S, Rigazio L. Towards deep neural network architectures robust to adversarial examples. arXiv preprint arXiv:1412.5068. 2014.

[A82] Akhtar N, Liu J, Mian A. Defense against universal adversarial perturbations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2018. 3389-3398.

[A83] Barreno M, Nelson B, Sears R, Joseph AD, Tygar JD. Can machine learning be secure? In: Proceedings of the 2006 ACM Symposium on Information, Computer and Communications Security. ACM, 2006. 16-25.

[A84] Barreno M, Nelson B, Joseph AD, Tygar JD. The security of machine learning. Machine Learning. 2010,81(2):121-148.

[A85] Papernot N, McDaniel P, Goodfellow I, Jha S, Celik ZB, Swami A. Practical black-box attacks against machine learning. In: Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security. ACM, 2017. 506-519.

[A86] Athalye A, Carlini N, Wagner D. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In: Internatioanl Conference on Machine Learning. PMLR, 2018. 274-283.

[A87] He W, Wei J, Chen X, Carlini N, Song D. Adversarial example defense: Ensembles of weak defenses are not strong. In: 11th USENIX Workshop on Offensive Technologies (WOOT 2017). 2017.

[A88] Liu X, Cheng M, Zhang H, Hsieh CJ. Towards robust neural networks via random self-ensemble. In: Proceedings of the European Conference on Computer Vision (ECCV). 2018. 369–385.

[A89] Corona I, Biggio B, Contini M, Piras L, Corda R, Mereu M, Mureddu G, Ariu D, Roli F. Deltaphish: Detecting phishing webpages in compromised websites. In: European Symposium on Research in Computer Security. Springer, 2017. 370–388.

[A90] Biggio B, Corona I, He ZM, Chan PP, Giacinto G, Yeung DS, Roli F. One-and-a-half-class multiple classifier systems for secure learning against evasion attacks at test time. In: International Workshop on Multiple Classifier Systems. Springer, 2015. 168-180.

[A91] Liu X, Li Y, Wu C, Hsieh CJ. Adv-bnn: Improved adversarial defense through robust Bayesian neural network. arXiv preprint arXiv: 1810.01279. 2018.

[A92] Lecuyer M, Atlidakis V, Geambasu R, Hsu D, Jana S. Certified robustness to adversarial examples with differential privacy. In: 2019 IEEE Symposium on Security and Privacy. IEEE, 2019. 656-572.

[A93] Gebru T, Morgenstern J, Vecchione B, Vaughan JW, Wallach H, Lii HD, Crawford K. Datasheets for datasets. Communications of the ACM. 2021,64(12):86-92

[A94] Hohman F, Wongsuphasawat K, Kery MB, Patel K. Understanding and visualizing data iteration in machine learning. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. ACM, 2020. 1-13.

[A95] Holland S, Hosny A, Newman S, Joseph J, Chemielinski K. The dataset nutrition label: A framework to drive higher data quality standards.arXiv preprint arXiv: 1805.03677. 2018.

[A96] Shi L, Wei F, Liu S, Tan L, Lian X, Zhou MX. Understanding text corpora with multiple facets. In: 2010 IEEE Symposium on Visual Analytics Science and Technology. IEEE, 2010. 99-106.

[A97] Smilkov D, Thorat N, Nicholson C, Reif E, Viégas FB, Wattenberg M. Embedding projector: Interactive visualization and interpretation of embeddings. arXiv preprint arXiv:1611.05469. 2016.

[A98] Chen T, Guestrin C. Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2016. 785-794.

[A99] Haufe S, Meinecke F, Görgen K, Dähne S, Haynes JD, Blankertz B, Bießmann F. On the interpretation of weight vectors of linear models in multivariate neuroimaging. Neuroimage. 2014,87:96-110.

[A100] Cano A, Zafra A, Ventura S. An interpretable classification rule mining algorithm. Information Sciences. 2013, 240:1-20.

[A101] Lakkaraju H, Bach SH, Leskovec J. Interpretable decision sets: A joint framework for description and prediction. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2016. 1675-1684

[A102] Wang J, Fujimaki R, Motohashi Y. Trading interpretability for accuracy: Oblique treed sparse additive models. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2015. 1245-1254.

[A103] Bien J, Tibshirani R. Prototype selection for interpretable classification. The Annals of Applied Statistics. 2011,5(4):2403-2424.

[A104] Endriss U, Leite J. Multi-objective learning of accurate and comprehensible classifiers–a case study. In: STAIRS 2014: Proceedings of the 7th European Starting AI Researcher Symposium. IOS Press, 2014. 264: 220.

[A105] Ustun B, Rudin C. Supersparse linear integer models for optimized medical scoring systems. Machine Learning. 2016,102(3):349-391.

[A106] Dash S, Gunluk O, Wei D. Boolean decision rules via column generation. Advances in Neural Information Processing Systems. 2018,31:4655-4665.

[A107] Wei D, Dash S, Gao T, Gunluk O. Generalized linear rule models. In: International Conference on Machine Learning. PMLR, 2019. 6687-6696.

[A108] Hind M, Wei D, Campbell M, Codella NC, Dhurandhar A, Mojsilović A, Ramamurthy K N, Varshney KR. TED: Teaching AI to explain its decisions. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. ACM, 2019. 123-129.

[A109] Lou Y, Caruana R, Gehrke J. Intelligible models for classification and regression. In: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2012. 150-158.

[A110] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473. 2014.

[A111] Xu K, Ba J, Kiros R, Cho K, Courville A, Salakhudinov R, Zemel R, Bengio Y. Show, attend and tell: Neural image caption generation with visual attention. In: International Conference on Machine Learning. PMLR, 2015. 2048-2057.

[A112] Yang Z, Yang D, Dyer C, He X, Smola A, Hovy E. Hierarchical attention networks for document classification. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2016. 1480-1489.

[A113] He X, He Z, Song J, Liu Z, Jiang YG, Chua TS. Nais: Neural attentive item similarity model for recommendation. IEEE

Transactions on Knowledge and Data Engineering. 2018,30(12):2354-2366.

[A114] Ying H, Zhuang F, Zhang F, Zhang F, Liu Y, Xu G, Xie X, Xiong H, Wu J. Sequential recommender system based on hierarchical attention network. In: IJCAI International Joint Conference on Artificial Intelligence. 2018. 3926-3932.

[A115] Zhou C, Bai J, Song J, Liu X, Zhao Z, Chen X, Gao J. Atrank: An attention-based user behavior modeling framework for recommendation. In: Thirty-Second AAAI Conference on Artificial Intelligence. AAAI, 2018. 4564-4571.

[A116] Yu S, Wang Y, Yang M, Li B, Qu Q, Shen J. NAIRS: A neural attentive interpretable recommendation system. In: Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining. ACM, 2019. 790-793.

[A117] Seo S, Huang J, Yang H, Liu Y. Interpretable convolutional neural networks with dual local and global attention for review rating prediction. In: Proceedings of the eleventh ACM conference on recommender systems. ACM, 2017. 297-305.

[A118] Kim B, Khanna R, Koyejo OO. Examples are not enough, learn to criticize! criticism for interpretability. Advances in Neural Information Processing Systems. 2016,29:2288-2296.

[A119] Kim B, Rudin C, Shah JA. The bayesian case model: A generative approach for case-based reasoning and prototype classification. Advances in Neural Information Processing Systems. 2014,27:1952-1960.

[A120] Mothilal RK, Sharma A, Tan C. Explaining machine learning classifiers through diverse counterfactual explanations. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. ACM, 2020. 607-617.

[A121] Wachter S, Mittelstadt B, Russell C. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Harvard Joural of Law & Technology. 2017, 31: 841-887.

[A122] Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531. 2015.

[A123] Kim B, Wattenberg M, Gilmer J, Cai C, Wexler J, Viegas F. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In: International Conference on Machine Learning. PMLR, 2018. 2668-2677

[A124] Xu K, Park DH, Yi C, Sutton C. Interpreting deep classifiers by visual distillation of dark knowledge. arXiv preprint arXiv:1803.04042. 2018.

[A125] Yang C, Rangarajan A, Ranka S. Global model interpretation via recursive partitioning. In: 2018 IEEE 20th International Conference on High Performance Computing and Communications; IEEE 16th International Conference on Smart City; IEEE 4th International Conference on Data Science and Systems (HPCC/SmartCity/DSS). IEEE, 2018. 1563-1570.

[A126] Zarlenga ME, Shams Z, Jamnik M. Efficient decompositional rule extraction for deep neural networks. arXiv preprint arXiv:2111.12628.

[A127] Bondarenko A, Zmanovska T, Borisov A. Decompositional rules extraction methods from neural networks. In: Proceedings of the 16th International Conference on Soft Computing MENDEL'10. 2010. 256-262.

[A128] Biswas SK, Chakraborty M, Purkayastha B, Roy P, Thounaojam DM. Rule extraction from training data using neural network. International Journal on Artificial Intelligence Tools. 2017,26(03):175006.

[A129] Zhou ZH, Jiang Y, Chen SF. Extracting symbolic rules from trained neural network ensembles. Ai Communications. 2003,16(1):3-15.

[A130] De Fortuny E J, Martens D. Active learning-based pedagogical rule extraction. IEEE transactions on neural networks and learning systems. 2015,26(11):2664-2677.

[A131] Casalicchio G, Molnar C, Bischl B. Visualizing the feature importance for black box models. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, 2018. 655-670.

[A132] Welling SH, Refsgaard HH, Brockhoff PB, Clemmensen LH. Forest floor visualizations of random forest. arXiv preprint arXiv:1605.09196. 2016.

[A133] Zhou B, Sun Y, Bau D, Torralba A. Interpretable basis decomposition for visual explanation. In: Proceedings of the European Conference on Computer Vision (ECCV). 2018. 119-134.

[A134] Baehrens D, Schroeter T, Harmeling S, Kawanabe M, Hansen K, Müller KR. How to explain individual classification decisions. The Journal of Machine Learning Research. 2010,11:1803-1831.

[A135] Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision. 2017. 618-626.

[A136] Zhang Q, Wu YN, Zhu SC. Interpretable convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 8827-8836.

[A137] Greenwell BM. pdp: An R package for constructing partial dependence plots. R Journal. 2017,9(1):421-436.

[A138] Mahendran A, Vedaldi A. Understanding deep image representations by inverting them. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015. 5188-5196.

[A139] Erhan D, Bengio Y, Courville A, Vincent P. Visualizing higher-layer features of a deep network. University of Montreal. 2019,1341(3), 1.

[A140] Zeiler MD, Fergus R. Visualizing and understanding convolutional networks. In: European Conference on Computer Vision. Springer, 2014. 818-833.

[A141] Karpathy A, Johnson J, Fei-Fei L. Visualizing and understanding recurrent networks. arXiv preprint arXiv: 1506.02078. 2015.

[A142] Craven M, Shavlik J. Extracting tree-structured representations of trained networks. Advances in neural information processing systems. 1995, 8.

[A143] Frosst N, Hinton G. Distilling a neural network into a soft decision tree. arXiv preprint arXiv:1711.09784. 2017.

[A144] Nguyen A, Dosovitskiy A, Yosinski J, Brox T, Clune J. Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. Advances in Neural Information Processing Systems. 2016,29.

[A145] Ribeiro MT, Singh S, Guestrin C. "Why should i trust you?" Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2016. 1135-1144.

[A146] Zilke JR, Loza Mencía E, Janssen F. Deepred–rule extraction from deep neural networks. In: International Conference on Discovery Science. Springer, 2016. 457-473.

[A147] Fong RC, Vedaldi A. Interpretable explanations of black boxes by meaningful perturbation. In: Proceedings of the IEEE International Conference on Computer Vision. 2017. 3429-3437.

[A148] Shrikumar A, Greenside P, Kundaje A. Learning important features through propagating activation differences. In: International Conference on Machine Learning. PMLR, 2017. 3145-3153

[A149] Simonyan K, Vedaldi A, Zisserman A. Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034. 2013.

[A150] Sundararajan M, Taly A, Yan Q. Gradients of counterfactuals. arXiv preprint arXiv:1611.02639. 2016.

[A151] Bach S, Binder A, Montavon G, Klauschen F, Muller KR, Samek W. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PloS one. 2015,10(7):e0130140.

[A152] Binder A, Montavon G, Lapuschkin S, Müller KR, Samek W. Layer-wise relevance propagation for neural networks with local renormalization layers. In: International Conference on Artificial Neural Networks. Springer, 2016. 63-71.

[A153] Binder A, Bach S, Montavon G, Müller KR, Samek W. Layer-wise relevance propagation for deep neural network architecture. In: Information Science and Applications (ICISA). Springer, 2016. 913-922.

[A154] Arachchige PCM, Bertok P, Khalil I, Liu D, Camtepe S, Atiquzzaman M. Local differential privacy for deep learning. IEEE Internet of Things Journal. 2019,7(7):5827-5842.

[A155] Yang R, Ma X, Bai X, Su X. Differential privacy images protection based on generative adversarial network. In: 2020 IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom). IEEE, 2020. 1688-1695.

[A156] Erlingsson Ú, Pihur V, Korolova A. Rappor: Randomized aggregatable privacy-preserving ordinal response. In: Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security. ACM, 2014. 1054-1067.

[A157] Beaulieu-Jones BK, Wu ZS, Williams C, Lee R, Bhavnani SP, Byrd JB, Greene CS. Privacy-preserving generative deep neural networks support clinical data sharing. Circulation: Cardiovascular Quality and Outcomes. 2019,12(7):e005122.

[A158] Xie L, Lin K, Wang S, Wang F, Zhou J. Differentially private generative adversarial network. arXiv preprint arXiv:1802.06739. 2018.

[A159] Bindschaedler V, Shokri R, Gunter CA. Plausible deniability for privacy-preserving data synthesis. arXiv preprint arXiv:1708.07975. 2017.

[A160] Kifer D, Lin BR. Towards an axiomatization of statistical privacy and utility. In: Proceedings of the twenty-ninth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems. ACM, 2010. 147-158.

[A161] Ye QQ, Meng XF, Zhu MJ, Huo Z. Survey on local differential privacy. Ruan Jian Xue Bao/Journal of Software. 2018, 29(7): 1981–2005 (in Chinese with English abstract).

[A162] Phan NH, Vu M, Liu Y, Jin R, Dou D, Wu X, Thai MT. Heterogeneous Gaussian mechanism: Preserving differential privacy in deep learning with provable robustness. arXiv preprint arXiv:1906.01444. 2019.

[A163] Phan NH, Wu X, Hu H, Dou D. Adaptive laplace mechanism: Differential privacy preservation in deep learning. In: 2017 IEEE International Conference on Data Mining (ICDM). IEEE, 2017. 385-394.

[A164] Shokri R, Shmatikov V. Privacy-preserving deep learning. In: Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security. ACM, 2015. 1310-1321.

[A165] Liu M, Jiang H, Chen J, Badokhon A, Wei X, Huang MC. A collaborative privacy-preserving deep learning system in distributed mobile environment. In: 2016 International Conference on Computational Science and Computational Intelligence (CSCI). IEEE, 2016. 192-197.

[A166] Song S, Chaudhuri K, Sarwate AD. Stochastic gradient descent with differentially private updates. In: 2013 IEEE Global Conference on Signal and Information Processing. IEEE, 2013. 245-248.

[A167] Wang N, Xiao X, Yang Y, Zhao J, Hui SC, Shin H, Shin J, Yu G. Collecting and analyzing multidimensional data with local differential privacy. In: 2019 IEEE 35th International Conference on Data Engineering (ICDE). IEEE, 2019. 638-649.

[A168] McMahan HB, Ramage D, Talwar K, Zhang L. Learning differentially private recurrent language models. arXiv preprint arXiv:1710.06963. 2017.

[A169] Wang Y, Si C, Wu X. Regression model fitting under differential privacy and model inversion attack. In: Twenty-fourth International Joint Conference on Artificial Intelligence. AAAI Press, 2015. 1003-1009.

[A170] Phan NH, Wang Y, Wu X, Dou D. Differential privacy preservation for deep auto-encoders: an application of human behavior prediction. In: Thirtieth AAAI Conference on Artificial Intelligence. AAAI Press, 2016. 1309-1316.

[A171] Phan NH, Wu X, Dou D. Preserving differential privacy in convolutional deep belief networks. Machine learning. 2017,106(9):1681-1704.

[A172] Zhang J, Zhang Z, Xiao X, Yang Y, Winslett M. Functional mechanism: regression analysis under differential privacy. arXiv preprint arXiv:1208.0219. 2012.

[A173] Chaudhuri K, Monteleoni C, Sarwate AD. Differentially private empirical risk minimization. Journal of Machine Learning Research. 2011,12(3):1069–1109.

[A174] Chaudhuri K, Monteleoni C. Privacy-preserving logistic regression. Advances in Neural Information Processing Systems. 2008,21:289–296.

[A175] Adesuyi TA, Kim BM. A layer-wise perturbation based privacy preserving deep neural networks. In: 2019 International Conference on Artificial Intelligence in Information and Communication (ICAIIC). IEEE, 2019. 389-394.

[A176] Acs G, Melis L, Castelluccia C, De Cristofaro E. Differentially private mixture of generative neural networks. IEEE Transactions on Knowledge and Data Engineering. 2018,31(6):1109-1121.

[A177] Zhang J, Zheng K, Mou W, Wang L. Efficient private ERM for smooth objectives. arXiv preprint arXiv:1703.09947. 2017.

[A178] Yuan D, Zhu X, Wei M, Ma J. Collaborative deep learning for medical image analysis with differential privacy. In: 2019 IEEE Global Communications Conference (GLOBECOM). IEEE, 2019. 1-6.

[A179] Rubinstein BIP, Bartlett PL, Huang L, Taft N. Learning in a large function space: Privacy-preserving mechanisms for SVM learning. arXiv preprint arXiv:0911.5708. 2009.

[A180] Papernot N, Abadi M, Erlingsson U, Goodfellow I, Talwar K. Semi-supervised knowledge transfer for deep learning from private training data. arXiv preprint arXiv:1610.05755. 2016.

[A181] Jayaraman B, Wang L, Evans D, Gu Q. Distributed learning without distress: Privacy-preserving empirical risk minimization.

Advances in Neural Information Processing Systems. 2018, 31.

[A182] Papernot N, Song S, Mironov I, Raghunathan A, Talwar K, Erlingsson U. Scalable private learning with pate. arXiv preprint arXiv:1802.08908. 2018.

[A183] Wu X, Li F, Kumar A, Chaudhuri L, Jha S, Naughton J. Bolt-on differential privacy for scalable stochastic gradient descent-based analytics. In: Proceedings of the 2017 ACM International Conference on Management of Data. ACM, 2017. 1307-1322.

[A184] Bost B, Popa RA, Tu S, Goldwasser S. Machine learning classification over encrypted data. Cryptology ePrint Archive. 2014.

[A185] Brickell J, Porter DE, Shmatikov V, Witchel E. Privacy-preserving remote diagnostics. In: Proceedings of the 14th ACM Conference on Computer and Communications Security. ACM, 2007. 498-507.

[A186] Gilad-Bachrach R, Dowlin N, Laine K, Lauter K, Naehrig M, Wernsing J. Cryptonets: Applying neural networks to encrypted data with high throughput and accuracy. In: International Conference on Machine Learning. PMLR, 2016. 201–210.

[A187] Graepel T, Lauter K, Naehrig M. ML confidential: Machine learning on encrypted data. In: International Conference on Information Security and Cryptology. Springer, 2012. 1-21.

[A188] Hesamifard E, Takabi H, Ghasemi M, Wright RN. Privacy-preserving machine learning as a service. In: Proceedings of the Conference on Privacy Enhancing Technologies (PETS'19). 2018. 123–142.

[A189] Li P, Li J, Huang Z, Li T, Gao C Z, Yiu SM, Chen K. Multi-key privacy-preserving deep learning in cloud computing. Future Generation Computer Systems. 2017,74:76-85.

[A190] Hesamifard E, Takabi H, Ghasemi M. Cryptodl: Deep neural networks over encrypted data. arXiv preprint arXiv:1711.05189. 2017.

[A191] Barni M, Orlandi C, Piva A. A privacy-preserving protocol for neural-network-based computation. In: Proceedings of the 8th Workshop on Multimedia and Security. ACM, 2006. 146-151.

[A192] Orlandi C, Piva A, Barni M. Oblivious neural network computing via homomorphic encryption. EURASIP Journal on Information Security. 2007:1-11.

[A193] Aono Y, Hayashi T, Wang L, Moriai S. Privacy-preserving deep learning via additively homomorphic encryption. IEEE Transactions on Information Forensics and Security. 2017,13(5):1333-1345.

[A194] Damgård I, Pastro V, Smart N, Zalarias S. Multiparty computation from somewhat homomorphic encryption. In: Annual Cryptology Conference. Springer, 2012. 643-662.

[A195] Chabanne H, De Wargny A, Milgram J, Morel C, Prouff E. Privacy-preserving classification on deep neural network. Cryptology ePrint Archive. 2017.

[A196] Mohassel P, Zhang Y. Secureml: A system for scalable privacy-preserving machine learning. In: 2017 IEEE Symposium on Security and Privacy (SP). IEEE, 2017. 19-38.

[A197] Mehnaz S, Bellala G, Bertino E. A secure sum protocol and its application to privacy-preserving multi-party analytics. In: Proceedings of the 22nd ACM on Symposium on Access Control Models and Technologies. ACM, 2017. 219-230.

[A198] Makri E, Rotaru D, Smart NP, Vercauteren F. EPIC: efficient private image classification (or: Learning from the masters). In: Cryptographers' Track at the RSA Conference. Springer, 2019. 473-492.

[A199] Bonawitz K, Ivanov V, Kreuter B, Marcedone A, McMahan B, Patel S, Rammage D, Segal A, Seth K. Practical secure aggregation for privacy-preserving machine learning. In: Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security. ACM, 2017. 1175-1191.

[A200] Danner G, Jelasity M. Fully distributed privacy preserving mini-batch gradient descent learning. In: IFIP International Conference on Distributed Applications and Interoperable Systems. Springer, 2015. 30-44.

[A201] McMahan B, Moore E, Ramage D, Hampson S, y Arcas BA. Communication-efficient learning of deep networks from decentralized data. In: Artificial Intelligence and Statistics. PMLR, 2017. 1273-1282.

[A202] Agrawal N, Shahin Shamsabadi A, Kusner MJ, Gascón A. QUOTIENT: two-party secure neural network training and prediction. In: Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security. ACM, 2019. 1231-1247.

[A203] Mohassel P, Rindal P. ABY3: A mixed protocol framework for machine learning. In: Proceedings of the 2018 ACM SIGSAC

Conference on Computer and Communications Security. ACM, 2018. 35-52.

[A204] Liu J, Juuti M, Lu Y, Asokan N. Oblivious neural network predictions via minionn transformations. In: Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security. ACM, 2017. 619-631.

[A205] Riazi MS, Weinert C, Tkachenko O, Songhori EM, Schneider T, Koushanfar F. Chameleon: A hybrid secure computation framework for machine learning applications. In: Proceedings of the 2018 on Asia conference on computer and communications security. ACM, 2018. 707-721.

[A206] Bunn P, Ostrovsky R. Secure two-party k-means clustering. In: Proceedings of the 14th ACM Conference on Computer and Communications Security. ACM, 2007. 486-497.

[A207] Jagannathan G, Wright RN. Privacy-preserving distributed k-means clustering over arbitrarily partitioned data. In: Proceedings of the eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining. ACM, 2005. 593-599.

[A208] Huang K, Liu X, Fu S, Guo D, Xu M. A lightweight privacy-preserving CNN feature extraction framework for mobile sensing. IEEE Transactions on Dependable and Secure Computing. 2019,18(3):1441-1455.

[A209] Jiang L, Tan R, Lou X, Lin G. On lightweight privacy-preserving collaborative learning for internet-of-things objects. In: Proceedings of the International Conference on Internet of Things Design and Implementation. ACM, 2019. 70-81.

[A210] Jia Q, Guo L, Jin Z, Fang Y. Preserving model privacy for machine learning in distributed systems. IEEE Transactions on Parallel and Distributed Systems. 2018,29(8):1808-1822.

[A211] Nikolaenko V, Weinsberg U, Ioannidis S, Joye M, Boneh D, Taft N. Privacy-preserving ridge regression on hundreds of millions of records. In: 2013 IEEE symposium on security and privacy. IEEE, 2013. 334-348.

[A212] Sanil AP, Karr AF, Lin X, Reiter JP. Privacy preserving regression modelling via distributed computation. In: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2004. 677-682.

[A213] Gascón A, Schoppmann P, Balle B, Raykova M, Doerner J, Zahur S, Evans D. Secure linear regression on vertically partitioned datasets. IACR Cryptol. eprint Arch. 2016: 892.

[A214] Slavkovic AB, Nardi Y, Tibbits MM. "Secure" Logistic Regression of Horizontally and Vertically Partitioned Distributed Databases. In: Seventh IEEE International Conference on Data Mining Workshops (ICDMW 2007). IEEE, 2007. 723-728.

[A215] Wu S, Teruya T, Kawamoto J, Sakuma J, Kikuchi H. Privacy-preservation for stochastic gradient descent application to secure logistic regression. In: The 27th annual conference of the Japanese Society for Artificial Intelligence. 2013. 27:1-4.

[A216] Garfinkel S. De-identification of Personal Information. US Department of Commerce, National Institute of Standards and Technology. 2015.

[A217] Khalil M, Ebner M. De-identification in learning analytics. Journal of Learning Analytics. 2016,3(1):129-138.

[A218] Dworkin M. Recommendation for block cipher modes of operation: methods for format-preserving encryption. NIST Special Publication. 2016, 800: 38G.

[A219] Kerschbaum F. Frequency-hiding order-preserving encryption. In: Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security. ACM, 2015. 656-667.

[A220] Stalla-Bourdillon S, Knight A. Anonymous data v. personal data-false debate: an EU perspective on anonymization, pseudonymization and personal data. Wis. Int'l LJ, 2016, 34: 284.

[A221] Youm HY. An overview of de-identification techniques and their standardization directions. IEICE TRANSACTIONS on Information and Systems. 2020,103(7):1448-1461.

[A222] Geyer RC, Klein T, Nabi M. Differentially private federated learning: A client level perspective. arXiv preprint arXiv:1712.07557. 2017.

[A223] Hao M, Li H, Luo X, Xu G, Yang H, Liu S. Efficient and privacy-enhanced federated learning for industrial artificial intelligence. IEEE Transactions on Industrial Informatics. 2019,16(10):6532-6542.

[A224] Konečný J, McMahan HB, Yu FX, Richtárik P, Suresh AT, Bacon D. Federated learning: Strategies for improving communication efficiency. arXiv preprint arXiv:1610.05492. 2016.

[A225] Hao M, Li H, Xu G, Liu S, Yang H. Towards efficient and privacy-preserving federated deep learning. In: ICC 2019-2019 IEEE

International Conference on Communications (ICC). IEEE, 2019. 1-6.

[A226] Hamm J. Minimax filter: Learning to preserve privacy from inference attacks. The Journal of Machine Learning Research. 2017,18(1):4704-4734.

[A227] Dufour-Sans E, Gay R, Pointcheval D. Reading in the dark: Classifying encrypted digits with functional encryption. Cryptology ePrint Archive. 2018.

[A228] Marc T, Stopar M, Hartman J, Bizjak M, Modic J. Privacy-enhanced machine learning with functional encryption. In: European Symposium on Research in Computer Security. Springer, 2019. 3-21.

[A229] Xu D, Yuan S, Zhang L, Wu X. FairGAN: Fairness-aware generative adversarial networks. In: 2018 IEEE International Conference on Big Data. IEEE, 2018. 570-575.

[A230] Kamiran F, Calders T. Data preprocessing techniques for classification without discrimination. Knowledge and Information Systems. 2012,33(1):1-33.

[A231] Backurs A, Indyk P, Onak K, Schieber B, Vakilian A, Wagner T. Scalable fair clustering. In: International Conference on Machine Learning. PMLR, 2019. 405-413.

[A232] Brunet ME, Alkalay-Houlihan C, Anderson A, Zemel R. Understanding the origins of bias in word embeddings. In: International Conference on Machine Learning. PMLR, 2019. 803-811.

[A233] Calmon F, Wei D, Vinzamuri B, Natesan Ramamurthy K, Varshney KR. Optimized pre-processing for discrimination prevention. Advances in Neural Information Processing Systems. 2017,30:3992-4001.

[A234] Luong BT, Ruggieri S, Turini F. 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. ACM, 2011. 502-510.

[A235] Mehrabi N, Morstatter F, Peng N, Galstyan A. Debiasing community detection: The importance of lowly connected nodes. In: 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). IEEE, 2019. 509-512.

[A236] Ruggieri S. Using t-closeness anonymity to control for non-discrimination. Transactions on Data Privacy. 2014,7(2): 99-129.

[A237] Sablayrolles A, Douze M, Schmid C, Jégou H. Radioactive data: tracing through training. In: International Conference on Machine Learning. PMLR, 2020. 8326-8335.

[A238] Samadi S, Tantipongpipat U, Morgenstern JH, Singh M, Vempala S. The price of fair pca: One extra dimension. Advances in neural information processing systems. 2018,31.

[A239] Madras D, Creager E, Pitassi T, Zemel R. Learning adversarially fair and transferable representations. In: International Conference on Machine Learning. PMLR, 2018. 3384-3393.

[A240] Zhao H, Coston A, Adel T, Gordon GJ. Conditional Learning of Fair Representations. arXiv preprint arXiv:1910.07162. 2019.

[A241] Xu D P, Yuan S, Zhang L, Wu X. FairGAN+: Achieving fair data generation and classification through generative adversarial nets. In: 2019 IEEE Conference on Big Data. IEEE, 2019. 1401-1406.

[A242] Xu D, Wu Y, Yuan S, Zhang L, Wu X. Achieving causal fairness through generative adversarial networks. In: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence. 2019. 1452-1458

[A243] Creager E, Madras D, Jacobsen JH, Weis M, Swersky K, Pitassi T, Zemel R. Flexibly fair representation learning by disentanglement. In: International Conference on Learning Representations. PMLR, 2019. 1436-1445.

[A244] Lahoti P, Gummadi KP, Weikum G. Operationalizing individual fairness with pairwise fair representations. arXiv preprint arXiv:1907.01439. 2019.

[A245] Agarwal A, Lohia P, Nagar S, Dey K, Saha D. Automated test generation to detect individual discrimination in AI models. arXiv preprint arXiv: 1809.03260. 2018.

[A246] Celis LE, Deshpande A, Kathuria T, Vishnoi NK. How to be fair and diverse? arXiv preprint arXiv: 1610.07183. 2016.

[A247] Kocaoglu M, Snyder C, Dimakis AG, Vishwanath S. Causalgan: Learning causal implicit generative models with adversarial training. arXiv preprint arXiv: 1709.02023. 2017.

[A248] Gordaliza P, Del Barrio E, Fabrice G, Loubes JM. Obtaining fairness using optimal transport theory. In: International Conference on Machine Learning. PMLR, 2019. 2357-2365.

[A249] Zemel R, Wu Y, Swersky K, Pitassi T, Dwork C. Learning fair representations. In: International Conference on Machine Learning. PMLR, 2013. 325-333.

Berk R, Heidari H, Jabbari S, Joseph M, Kearns M, Morgenstern J, Neel S, Roth A. A convex framework for fair regression. arXiv preprint arXiv:1706.02409. 2017.

[A251] Lepri B, Oliver N, Letouzé E, Pentland A, Vinck P. Fair, transparent, and accountable algorithmic decision-making processes. Philosophy & Technology. 2018,31(4):611-627.

[A252] Bechavod Y, Ligett K. Penalizing unfairness in binary classification. arXiv preprint arXiv:1707.00044. 2017.

[A253] Huang L, Vishnoi N. Stable and fair classification. In: International Conference on Machine Learning. PMLR, 2019. 2879-2890.

[A254] Kamiran F, Calders T, Pechenizkiy M. Discrimination aware decision tree learning. In: 2010 IEEE International Conference on Data Mining. IEEE, 2010. 869-874.

[A255] Kamishima T, Akaho S, Asoh H, Sakuma J. Fairness-aware classifier with prejudice remover regularizer. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, 2012. 35-50.

[A256] Quadrianto N, Sharmanska V. Recycling privileged learning and distribution matching for fairness. Advances in Neural Information Processing Systems. 2017,30:677-688.

[A257] Zafar MB, Valera I, Rogriguez MG, Gummadi KP. Fairness constraints: Mechanisms for fair classification. In: Artificial intelligence and statistics. PMLR, 2017. 962-970.

[A258] Zhang BH, Lemoine B, Mitchell M. Mitigating unwanted biases with adversarial learning. In: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society. ACM, 2018. 335-340.

[A259] Kusner MJ, Loftus J, Russell C, Silva R. Counterfactual fairness. Advances in neural information processing systems. 2017,30:4066-4076.

[A260] Bechavod Y, Ligett K. Learning fair classifiers: A regularization-inspired approach. arXiv preprint arXiv: 1707.00044v2. 2017.

[A261] Russell C, Kusner MJ, Loftus J, Silva R. When worlds collide: integrating different counterfactual assumptions in fairness. Advances in neural information processing systems. 2017,30:6414-6423.

[A262] Wu Y, Zhang L, Wu X, Tong H. Pc-fairness: A unified framework for measuring causality-based fairness. Advances in Neural Information Processing Systems. 2019,32:3399-3409.

[A263] Wu Y, Zhang L, Wu X. Counterfactual fairness: Unidentification, bound and algorithm. In: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence. 2019. 1438-1444.

[A264] Agarwal A, Dudík M, Wu ZS. Fair regression: Quantitative definitions and reduction-based algorithms. In: International Conference on Machine Learning. PMLR, 2019. 120-129.

[A265] Conitzer V, Freeman R, Shah N, Vaughan JW. Group fairness for the allocation of indivisible goods. In: Proceedings of the AAAI Conference on Artificial Intelligence. 2019. 33(01):1853-1860.

[A266] Kusner M, Russell C, Loftus J, Silva R. Making decisions that reduce discriminatory impacts. In: International Conference on Machine Learning. PMLR, 2019. 3591-3600.

[A267] Ustun B, Liu Y, Parkes D. Fairness without harm: Decoupled classifiers with preference guarantees. In: International Conference on Machine Learning. PMLR, 2019. 6373-6382.

[A268] Tsang A, Wilder B, Rice E, Tamble M, Zick Y. Group-fairness in influence maximization. arXiv preprint arXiv:1903.00967. 2019.

[A269] Chen X, Fain B, Lyu L, Munagala K. Proportionally fair clustering. In: International Conference on Machine Learning. PMLR, 2019. 1032-1041.

[A270] Agarwal A, Beygelzimer A, Dudík M, Langford J, Wallach H. A reductions approach to fair classification. In: International Conference on Machine Learning. PMLR, 2018. 60-69.

[A271] Goh G, Cotter A, Gupta M, Friedlander MP. Satisfying real-world goals with dataset constraints. Advances in Neural Information Processing Systems. 2016,29:2415-2423.

[A272] Grgić-Hlača N, Zafar MB, Gummadi KP, Weller A. Beyond distributive fairness in algorithmic decision making: Feature selection for procedurally fair learning. Proceedings of the AAAI Conference on Artificial Intelligence. 2018. 32(1).

[A273] Jung C, Kearns MJ, Neel S, Roth A, Stapleton L, Wu ZS. An algorithmic framework for fairness elicitation. arXiv preprint arXiv: 1905.10660. 2019.

[A274] Kim MP, Korolova A, Rothblum GN, Yona G. Preference-informed fairness. arXiv preprint arXiv: 1904.01793. 2019.

[A275] Zafar MB, Valera I, Gomez Rodriguez M, Gummadi KP. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In: Proceedings of the 26th International Conference on World Wide Web. ACM, 2017. 1171-1180.

[A276] Zafar MB, Valera I, Rodriguez M, Gummadi K, Weller A. From parity to preference-based notions of fairness in classification. Advances in Neural Information Processing Systems. 2017,30:229-239.

[A277] Chiappa S. Path-specific counterfactual fairness. In: Proceedings of the AAAI Conference on Artificial Intelligence. 2019,33(01):7801-7808.

[A278] Woodworth B, Gunasekar S, Ohannessian MI, Srebro N. Learning non-discriminatory predictors. In: Conference on Learning Theory. PMLR, 2017. 1920-1953.

[A279] Menon AK, Williamson RC. The cost of fairness in binary classification. In: Conference on Fairness, Accountability and Transparency. PMLR, 2018. 107-118.

[A280] Dwork C, Immorlica N, Kalai AT, Leiserson M. Decoupled classifiers for group-fair and efficient machine learning. In: Conference on Fairness, Accountability and Transparency. PMLR, 2018. 119-133.

[A281] Bolukbasi T, Chang KW, Zou JY, Saligrama V, Kalai AT. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. Advances in neural information processing systems. 2016,29:4349-4357.

[A282] Corbett-Davies S, Pierson E, Feller A, Goel S, Huq A. Algorithmic decision making and the cost of fairness. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2017: 797-806.

[A283] Jagielski M, Kearns M, Mao J, Oprea A, Roth A, Sharifi-Malvajerdi S, Ullman J. Differentially private fair learning. In: International Conference on Machine Learning. PMLR, 2019. 3000-3008.

[A284] Lohia PK, Ramamurthy KN, Bhide M, Saha D, Varshney KR, Puri R. Bias mitigation post-processing for individual and group fairness. Icassp 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (icassp). IEEE, 2019. 2847-2851.

[A285] McQuillan D. People's councils for ethical machine learning. Social Media + Society. 2018, 4(2): 2056305118768303.

[A286] Broeders D, Schrijvers E, van der Sloot B, Van Brakel R, de Hoog J, Ballin EH. Big Data and security policies: Towards a framework for regulating the phases of analytics and use of Big Data. Computer Law & Security Review. 2017,33(3):309-323.

[A287] Brennan-Marquez K. Plausible cause: Explanatory standards in the age of powerful machines. Vanderbilt Law Review. 2017,70:1249.

[A288] Bunt A, Lount M, Lauzon C. Are explanations always important? A study of deployed, low-cost intelligent interactive systems. In: Proceedings of the 2012 ACM international conference on Intelligent User Interfaces. ACM, 2012. 169-178.

[A289] Dobson JE. Can an algorithm be disturbed? Machine learning, intrinsic criticism, and the Digital Humanities. College Literature. 2015: 543-564.

[A290] Martin K. Ethical implications and accountability of algorithms. Journal of Business Ethics. 2019,160(4):835-850.

[A291] Pauleen DJ, Rooney D, Intezari A. Big data, little wisdom: trouble brewing? Ethical implications for the information systems discipline. Social Epistemology. 2017,31(4):400-416.

[A292] Feurer M, Klein A, Eggensperger K, Springenberg J, Blum M, Hutter F. Efficient and robust automated machine learning. Advances in Neural Information Processing Systems. 2015,28:2962-2970.

[A293] Stahl BC, Wright D. Ethics and privacy in AI and big data: Implementing responsible research and innovation. IEEE Security & Privacy. 2018,16(3):26-33.

[A294] Kroll JA. Accountable algorithms. University of Pennsylvania Law Review. 2015: 633-705.

[A295] Raji ID, Smart A, White RN, Mitchell M, Gebru T, Hutchinson B, Smith-Loud J, Theron D, Barnes P. Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In: Proceedings of the 2020 Conference on

Fairness, Accountability, and Transparency. ACM, 2020. 33-44.

[A296] LaBrie RC, Steinke G. Towards a framework for ethical audits of AI algorithms. In: Proceedings of the Conference on Data Science and Analytics for Decision Support. 2019.

[A297] Galdon Clavell G, Martín Zamorano M, Castillo C, Smith O, Matic A. Auditing algorithms: On lessons learned and the risks of data minimization. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. ACM, 2020. 265-271.

[A298] Rahwan I. Society-in-the-loop: programming the algorithmic social contract. Ethics and information technology. 2018,20(1):5-14.

[A299] Sandvig C, Hamilton K, Karahalios K, Langbort C. Auditing algorithms: Research methods for detecting discrimination on internet platforms. Data and Discrimination: Converting Critical Concerns into Productive Inquiry. 2014,22:4349-4357.

[A300] Veeramachaneni K, Arnaldo I, Korrapati V, Bassias C, Li K. AI^2: training a big data machine to defend. In: 2016 IEEE 2nd Conference on High Performance and Smart Computing (HPSC), and IEEE International Conference on Intelligent Data and Security (IDS). IEEE, 2016. 49-54.

[A301] Kaur D, Uslu S, Durresi A, Mohler G, Carter JG. Trust-based human-machine collaboration mechanism for predicting crimes. In: International Conference on Advanced Information Networking and Applications. Springer, 2020. 603-616.

[A302] Daugherty PR, Wilson HJ. Human + machine: reimaging work in the age of AI. Harvard Business Press. 2018.

[A303] Ruan Y, Zhang P, Alfantoukh L, Durresi A. Measurement theory-based trust management framework for online social communities. ACM Transactions on Internet Technology (TOIT). 2017,17(2):1-24.

[A304] Uslu S, Kaur D, Rivera SJ, Durresi A, Babbar-Sebens M, Tilt JH. Control theoretical modeling of trust-based decision making in food-energy-water management. In: Confference on Complex, Intelligent, and Software Intensive Systems. Springer, 2020. 97-107.

[A305] Uslu S, Kaur D, Rivera S J, Durresi A, Babbar-Sebens M. Trust-based decision making for food-energy-water actors. International Conference on Advanced Information Networking and Applications. Springer, 2020. 591-602.

[A306] Uslu S, Kaur D, Rivera SJ, Durresi A, Babbar-Sebens M. Trust-based game-theoretical decision making for food-energy-water management. In: International Conference on Broadband and Wireless Computing, Communication and Applications. Springer, 2019. 125-136.