Date	Topic	Title	Link
Jan 16	Fairness	Inherent Trade-Offs in the Fair Determination of Risk Scores	https://arxiv.org/abs/1609.05807
Jan 16	Fairness	The Possibility of Fairness: Revisiting the Impossibility Theorem in Practice	https://arxiv.org/abs/2302.06347
Jan 23	Ethics	On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?	https://dl.acm.org/doi/10.1145/3442188.3445922
Jan 23	Ethics	The Values Encoded in Machine Learning Research	https://arxiv.org/abs/2106.15590
Jan 23	Fairness	Stereotyping Norwegian Salmon: An Inventory of Pitfalls in Fairness Benchmark Datasets:	https://www.microsoft.com/en-us/research/uploads/prod/2021/06/The_Salmon_paper.pdf
Jan 30	Fairness	Image Cropping on Twitter: Fairness Metrics, their Limitations, and the Importance of Representation, Design, and Agency	https://arxiv.org/abs/2105.08667
Jan 30	Accountability	Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing	https://arxiv.org/abs/2001.00973
Jan 30	Fairness	Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness	https://arxiv.org/abs/1711.05144
Feb 6	Fairness	Feedback Loop and Bias Amplification in Recommender Systems	https://dl.acm.org/doi/pdf/10.1145/3340531.3412152

Feb 6	Fairness	Two-Sided Fairness for Repeated Matchings in Two-Sided Markets: A Case Study of a Ride-Hailing Platform	https://dl.acm.org/doi/10.1145/3292500.3330793
Feb 6	Fairness	Fairness is not static: deeper understanding of long term fairness via simulation studies	https://dl.acm.org/doi/abs/10.1145/3351095.3372878
Feb 20	Privacy	Membership Inference Attacks against Machine Learning Models	https://arxiv.org/abs/1610.05820
Feb 20	Privacy	The Privacy Onion Effect: Memorization is Relative	https://arxiv.org/abs/2206.10469
Feb 20	Privacy	Machine Unlearning	https://arxiv.org/abs/1912.03817
March 19	Robustness	On Evaluating Adversarial Robustness	https://arxiv.org/abs/1902.06705
March 19	Robustness	Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback	https://arxiv.org/abs/2307.15217
March 19	Safety	Jailbroken: How Does LLM Safety Training Fail?	https://arxiv.org/abs/2307.02483
March 26	Explainability	on the privacy risks of model explanations	https://arxiv.org/abs/1907.00164
March 26	Explainability	Counterfactual Explanations and Algorithmic Recourses for Machine Learning: A Review	https://arxiv.org/abs/2010.10596
March 26	Transparency/ Safety	Alignment of Language Agents	https://arxiv.org/abs/2103.14659