	PaperID	Title	Authors
Sunday, February 25, 9:30 - 10:45am			
Room 217	678	NeRFail: Neural Radiance Fields-based multiview adversarial attack	Wenxiang Jiang; Hanwei ZHANG; Xi WANG; Zhongwen Guo; Hao Wang
	10939	Robust Stochastic Graph Generator for Counterfactual Explanations	Mario Prado Romero; Bardh Prenkaj; Giovanni Stilo
	2365	Identifying Reasons for Bias: An Argumentation-Based Approach	Madeleine Waller; Odinaldo Rodrigues; Oana Cocarascu
	4159	Visual Adversarial Examples Jailbreak Aligned Large Language Models	Xiangyu Qi; Kaixuan Huang; Ashwinee Panda; Peter Henderson; Mengdi Wang; Prateek Mittal
	3147	Trade-Offs in Fine-Tuned Diffusion Models between Accuracy and Interpretability	Mischa Dombrowski; Hadrien Reynaud; Johanna Müller; Matthew Baugh; Bernhard Kainz
Room 211	3193	Pure-Past Action Masking	Giovanni Varricchione; Natasha Alechina; Mehdi Dastani; Giuseppe De Giacomo; Brian Logan; Giuseppe Perelli
	3672	MaxEnt Loss: Constrained Maximum Entropy for Calibration under Out-of-Distribution Shift	Dexter Neo; Stefan Winkler; Tsu Han Chen
	5369	Contrastive Credibility Propagation for Reliable Semi-supervised Learning	Brody Kutt; Pralay Ramteke; Xavier Mignot; Pamela Toman; Nandini Ramanan; Sujit Rokka Chhetri; Shan Huang; Min Du; William Hewlett
	6305	From Hope to Safety: Unlearning Biases of Deep Models via Gradient Penalization in Latent Space	Maximilian Dreyer; Frederik Pahde; Christopher Anders; Wojciech Samek; Sebastian Lapuschkin
Sunday, February 25, 11:15am - 12:30pm			
Room 217	7404	Balance Reward and Safety Optimization for Safe Reinforcement Learning: A Perspective of Gradient Manipulation	Shangding Gu; Bilgehan Sel; Yuhao Ding; Lu Wang; Qingwei Lin; Ming Jin; Alois Knoll
	7514	I Prefer Not to Say: Protecting User Consent in Models with Optional Personal Data	Tobias Leemann; Martin Pawelczyk; Christian Eberle; Gjergji Kasneci
	7575	Enumerating Safe Regions in Deep Neural Networks with Provable Probabilistic Guarantees	Luca Marzari; Davide Corsi; Enrico Marchesini; Alessandro Farinelli; Ferdinando Cicalese
	8061	Dissenting Explanations: Leveraging Disagreement to Reduce Model Overreliance	Omer Reingold; Judy Hanwen Shen; Aditi Talati
	8418	Towards Efficient Verification of Quantized Neural Networks	Pei Huang; Haoze Wu; Yuting Yang; Ieva Daukantas; Min Wu; Yedi Zhang; Clark Barrett
Sunday, February 25, 2:00 - 3:15pm			
Room 217	9040	PointCVaR: Risk-optimized Outlier Removal for Robust 3D Point Cloud Classification	Xinke Li; Junchi Lu; Henghui Ding; Changsheng Sun; Joey Tianyi Zhou; Yeow Meng Chee
	10629	GaLileo: General Linear Relaxation Framework for Tightening Robustness Certification of Transformers	Yun-Ruo Zhang; Lujia Shen; Shanqing GUO; Shouling Ji
	10738	Robust Uncertainty Quantification Using Conformalised Monte Carlo Prediction	Daniel Bethell; Simos Gerasimou; Radu Calinescu
	10183	Coevolutionary Algorithm for Building Robust Decision Trees under Minimax Regret	Adam Żychowski; Andrew Perrault; Jacek Mańdziuk
Room 211	12016	Stability Analysis of Switched Linear Systems with Neural Lyapunov Functions	Virginie Debauche; Alec Edwards; Raphaël Jungers; Alessandro Abate
	11524	Solving Non-rectangular Reward-Robust MDPs via Frequency Regularization	Uri Gadot; Esther Derman; Navdeep Kumar; Maxence Mohamed Elfatihi; Kfir Levy; Shie Mannor
	11952	A PAC Learning Algorithm for LTL and Omega-Regular Objectives in MDPs	Mateo Perez; Fabio Somenzi; Ashutosh Trivedi
	11978	Towards Fairer Centroids in K-means Clustering	Stanley Simoes; Deepak P; Muiris MacCarthaigh
	12398	Providing Fair Recourse over Plausible Groups	Jayanth Yetukuri; Ian Hardy; Yevgeniy Vorobeychik; Berk Ustun; Yang Liu
Sunday, February 25, 3:45 - 5:00pm			
Room 217	12441	On the Importance of Application-Grounded Experimental Design for Evaluating Explainable ML Methods	Kasun Amarasinghe; Kit Rodolfa; Sérgio Jesus; Valerie Chen; Vladimir Balayan; Pedro Saleiro; Pedro Bizarro; Ameet Talwalkar; Rayid Ghani
	12565	Learning Fair Policies for Multi-Stage Selection Problems from Observational Data	Zhuangzhuang Jia; Grani A Hanasusanto; Phebe Vayanos; Weijun Xie
	13716	Human-Guided Moral Decision Making in Text-Based Games	Zijing Shi; Meng Fang; Ling Chen; Yali Du; Jun WANG
	11313	Safe Reinforcement Learning with Instantaneous Constraints: The Role of Aggressive Exploration	Honghao Wei; Xin Liu; Lei Ying