



# AAAI-24: Safe, Robust and Responsible AI

| ID   | Title   | Authors  |
|------|---|--|
| 20   | Moderate Message Passing Improves Calibration: A Universal Way to Mitigate Confidence Bias in Graph Neural Networks | Min Wang; Hao Yang; Jincal Huang; Qing Cheng   |
| 408  | Self-Supervised Likelihood Estimation with Energy Guidance for Anomaly Segmentation in Urban Scenes                 | Yuanpeng Tu; Yuxi Li; Boshen Zhang; Liang Liu; Jiangning Zhang; Yabiao Wang; Cairong Zhao                  |
| 602  | Toward More Generalized Malicious URL Detection Models  | Yun-Da Tsai; Cayon Liow; Yin Sheng Siang; Shou-De Lin  |
| 678  | NeRFail: Neural Radiance Fields-Based Multiview Adversarial Attack  | Wenxiang Jiang; Hanwei ZHANG; Xi WANG; Zhongwen Guo; Hao Wang  |
| 1415 | Accelerating Adversarially Robust Model Selection for Deep Neural Networks via Racing                               | Matthias König; Holger Hoos; Jan Van Rijn  |
| 1421 | Would You Like Your Data to Be Trained? A User Controllable Recommendation Framework                                | Lei Wang; Xu Chen; Zhenhua Dong; Quanyu Dai  |
| 1513 | TTTS: Tree Test Time Simulation for Enhancing Decision Tree Robustness against Adversarial Examples                 | Seffi Cohen; Ofir Arbili; Yisroel Mirsky; Lior Rokach  |
| 1557 | Invisible Backdoor Attack against 3D Point Cloud Classifier in Graph Spectral Domain                                | Linkun Fan; Fazhi He; Tongzhen Si; Wei Tang; Bing Li   |
| 1587 | Conformal Prediction Regions for Time Series Using Linear Complementarity Programming                               | Matthew Cleaveland; Insup Lee; George Pappas; Lars   |
| 1647 | Toward Robustness in Multi-Label Classification: A Data Augmentation Strategy against Imbalance and Noise           | Hwanjun Song; Minseok Kim; Jae-Gil Lee   |
| 1739 | Revisiting the Information Capacity of Neural Network Watermarks: Upper Bound Estimation and Beyond                 | Fangqi Li; Haodong Zhao; Wei Du; Shilin Wang   |
| 2138 | Sparsity-Guided Holistic Explanation for LLMs with Interpretable Inference-Time Intervention                        | Zhen Tan; Tianlong Chen; Zhenyu Zhang; Huan Liu  |
| 2365 | Identifying Reasons for Bias: An Argumentation-Based Approach   | Madeleine Waller; Odinaldo Rodrigues; Oana Cocarascu   |
| 2418 | Generative Model for Decision Trees   | Riccardo Guidotti; Anna Monreale; Mattia Setzu; Giulia Volpi   |
| 2613 | Interpretability Benchmark for Evaluating Spatial Misalignment of Prototypical Parts Explanations                   | Mikołaj Sacha; Bartosz Jura; Dawid Rymarczyk; Łukasz Struski; Jacek Tabor; Bartosz Zieliński               |
| 2619 | Personalization as a Shortcut for Few-Shot Backdoor Attack against Text-to-Image Diffusion Models                   | Yihao Huang; Felix Juefei-Xu; Qing Guo; Jie Zhang; Yutong Wu; Ming Hu; Tianlin Li; Geguang Pu; Yang Liu    |
| 2787 | ORES: Open-Vocabulary Responsible Visual Synthesis  | Minheng Ni; Chenfei Wu; Xiaodong Wang; Shengming Yin; Lijuan Wang; Zicheng Liu; Nan Duan                   |
| 2816 | CCTR: Calibrating Trajectory Prediction for Uncertainty-Aware Motion Planning in Autonomous Driving                 | Chengtai Cao; Xinhong Chen; Jianping Wang; Qun Song; Rui Tan; Yung-Hui Li                                  |
| 2946 | Bidirectional Contrastive Split Learning for Visual Question Answering  | Yuwei Sun; Hideya Ochiai   |
| 3097 | Concealing Sensitive Samples against Gradient Leakage in Federated Learning   | Jing Wu; Munawar Hayat; Mingyi Zhou; Mehrtaash Harandi   |
| 3147 | Trade-Offs in Fine-Tuned Diffusion Models between Accuracy and Interpretability                                     | Mischa Dombrowski; Hadrien Reynaud; Johanna Müller; Matthew Baugh; Bernhard Kainz                          |
| 3193 | Pure-Past Action Masking  | Giovanni Varricchione; Natasha Alechina; Mehdi Dastani; Giuseppe De Giacomo; Brian Logan; Giuseppe Perelli |
| 3251 | Q-SENN: Quantized Self-Explaining Neural Networks   | Thomas Norrenbrock; Marco Rudolph; Bodo Rosenhahn  |
| 3308 | A Simple and Practical Method for Reducing the Disparate Impact of Differential Privacy                             | Lucas Rosenblatt; Julia Stoyanovich; Christopher Musco   |
| 3314 | Physics-Informed Representation and Learning: Control and Risk Quantification                                       | Zhuoyuan Wang; Reece Keller; Xiyu Deng; Kenta Hoshino; Takashi Tanaka; Yorie Nakahira                      |
| 3331 | SocialStigmaQA: A Benchmark to Uncover Stigma Amplification in Generative Language Models                           | Manish Nagireddy; Lamogha Chiazor; Moninder Singh; Ioana Baldini   |
| 3349 | Can LLM Replace Stack Overflow? A Study on Robustness and Reliability of Large Language Model Code Generation       | Li Zhong; Zilong Wang  |



|      |   |  |
|------|---|--|
| 3384 | Game-Theoretic Unlearnable Example Generator  | Shuang Liu; Yihan Wang; Xiao-Shan Gao  |
| 3535 | Truth Forest: Toward Multi-Scale Truthfulness in Large Language Models through Intervention without Tuning                              | Zhongzhi Chen; Xingwu Sun; Xianfeng Jiao; Fengzong Lian; Zhanhui Kang; Di Wang; Chengzhong Xu                                      |
| 3656 | A Framework for Data-Driven Explainability in Mathematical Optimization   | Kevin-Martin Aigner; Marc Goerigk; Michael Hartisch; Frauke Liers; Arthur Miehlch  |
| 3672 | MaxEnt Loss: Constrained Maximum Entropy for Calibration under Out-of-Distribution Shift  | Dexter Neo; Stefan Winkler; Tsu Han Chen   |
| 3715 | Exponent Relaxation of Polynomial Zonotopes and Its Applications in Formal Neural Network Verification                                  | Tobias Ladner; Matthias Althoff  |
| 3954 | Provable Robustness against a Union of $L_0$ Adversarial Attacks  | Zayd Hammoudeh; Daniel Lowd  |
| 4159 | Visual Adversarial Examples Jailbreak Aligned Large Language Models   | Xiangyu Qi; Kaixuan Huang; Ashwinee Panda; Peter Henderson; Mengdi Wang; Prateek Mittal  |
| 4347 | CASE: Exploiting Intra-class Compactness and Inter-class Separability of Feature Embeddings for Out-of-Distribution Detection           | Shuai Feng; Pengsheng Jin; Chongjun Wang   |
| 4354 | ImageCaptioner2: Image Captioner for Image Captioning Bias Amplification Assessment   | Eslam Abdelrahman; Pengzhan Sun; Li Erran Li; Mohamed Elhoseiny  |
| 4980 | I-CEE: Tailoring Explanations of Image Classifications Models to User Expertise   | Yao Rong; Peizhu Qian; Vaibhav Unhelkar; Enkelejda Kasneci   |
| 5103 | Reward Certification for Policy Smoothed Reinforcement Learning   | Ronghui Mu; Leandro Soriano Marcolino; Yanghao Zhang; Tianle Zhang; Xiaowei Huang; Wenjie Ruan                                     |
| 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 |
| 5414 | Constrained Meta-Reinforcement Learning for Adaptable Safety Guarantee with Differentiable Convex Programming                           | Minjae Cho; Chuangchuan Sun  |
| 5447 | Long-Term Safe Reinforcement Learning with Binary Feedback  | Akifumi Wachi; Wataru Hashimoto; Kazumune Hashimoto  |
| 5479 | DeepBern-Nets: Taming the Complexity of Certifying Neural Networks Using Bernstein Polynomial Activations and Precise Bound Propagation | Haitham Khedr; Yasser Shoukry  |
| 5530 | A Huber Loss Minimization Approach to Byzantine Robust Federated Learning   | Puning Zhao; Fei Yu; Zhiguo Wan  |
| 5731 | UMA: Facilitating Backdoor Scanning via Unlearning-Based Model Ablation   | Yue Zhao; Congyi Li; Kai Chen  |
| 5768 | EncryIP: A Practical Encryption-Based Framework for Model Intellectual Property Protection  | Xin Mu; Yu Wang; Zhengan Huang; Junzuo Lai; Yehong Zhang; Hui Wang; Yue Yu   |
| 6077 | Automatically Testing Functional Properties of Code Translation Models  | Hasan Ferit Eniser; Valentin Wüstholtz; Maria Christakis   |
| 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  |
| 6422 | Promoting Counterfactual Robustness through Diversity   | Francesco Leofante; Nico Potyka  |
| 6541 | $\pi$ -Light: Programmatic Interpretable Reinforcement Learning for Resource-Limited Traffic Signal Control                             | Yin Gu; Kai Zhang; Qi Liu; Weibo Gao; Longfei Li; Jun Zhou   |
| 6647 | On the Concept Trustworthiness in Concept Bottleneck Models   | Qihan Huang; Jie Song; Jingwen Hu; Haofei Zhang; Yong Wang; Mingli Song  |
| 6845 | Quantile-Based Maximum Likelihood Training for Outlier Detection  | Masoud Taghikhah; Nishant Kumar; Sinisa Segvic; Abouzar Eslami; Stefan Gumhold   |
| 7062 | Combining Graph Transformers Based Multi-Label Active Learning and Informative Data Augmentation for Chest Xray Classification          | Dwarikanath Mahapatra; Behzad Bozorgtabar; Zongyuan Ge; Mauricio Reyes; Jean-Philippe Thiran                                       |
| 7241 | P2BPO: Permeable Penalty Barrier-Based Policy Optimization for Safe RL  | Sumanta Dey; Pallab Dasgupta; Soumyajit Dey  |
| 7263 | Feature Unlearning for Pre-trained GANs and VAEs  | Saemi Moon; Seunghyuk Cho; Dongwoo Kim   |
| 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  |



|       |   |   |
|-------|---|---|
| 7418  | LR-XFL: Logical Reasoning-Based Explainable Federated Learning  | Yanci Zhang; Han Yu   |
| 7427  | Towards Large Certified Radius in Randomized Smoothing Using Quasiconcave Optimization                              | Bo-Han Kung; Shang-Tse Chen   |
| 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      |
| 7664  | Responsible Bandit Learning via Privacy-Protected Mean-Volatility Utility   | Shanshan Zhao; Wenhai Cui; Bei Jiang; Linglong Kong; Xiaodong Yan                             |
| 7936  | Efficient Toxic Content Detection by Bootstrapping and Distilling Large Language Models                             | Jiang Zhang; Qiong Wu; Yiming Xu; Cheng Cao; Zheng Du; Konstantinos Psounis                   |
| 7991  | Handling Long and Richly Constrained Tasks through Constrained Hierarchical Reinforcement Learning                  | Yuxiao LU; Arunesh Sinha; Pradeep Varakantham   |
| 8061  | Dissenting Explanations: Leveraging Disagreement to Reduce Model Overreliance                                       | Omer Reingold; Judy Hanwen Shen; Aditi Talati   |
| 8109  | Stronger and Transferable Node Injection Attacks  | Samyak Jain; Tanima Dutta   |
| 8248  | Representation-Based Robustness in Goal-Conditioned Reinforcement Learning  | Xiangyu Yin; Sihao Wu; Jiaxu Liu; Meng Fang; Xingyu Zhao; Xiaowei Huang; Wenjie Ruan          |
| 8287  | Analysis of Differentially Private Synthetic Data: A Measurement Error Approach                                     | Yangdi Jiang; Yi LIU; Xiaodong Yan; Anne-Sophie Charest; Linglong Kong; Bei Jiang             |
| 8418  | Towards Efficient Verification of Quantized Neural Networks   | Pei Huang; Haoze Wu; Yuting Yang; Ieva Daukantas; Min Wu; Yedi Zhang; Clark Barrett           |
| 8555  | Quilt: Robust Data Segment Selection against Concept Drifts   | Minsu Kim; Seong-Hyeon Hwang; Steven Euijong Whang  |
| 8658  | Understanding Likelihood of Normalizing Flow and Image Complexity through the Lens of Out-of-Distribution Detection | Genki Osada; Tsubasa Takahashi; Takashi Nishide   |
| 8677  | Byzantine-Robust Decentralized Learning via Remove-then-Clip Aggregation  | Caiyi Yang; Javad Ghaderi   |
| 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           |
| 9437  | Generating Diagnostic and Actionable Explanations for Fair Graph Neural Networks                                    | Zhenzhong Wang; Qingyuan Zeng; Wanyu LIN; Min Jiang; KC Tan                                   |
| 9612  | Rethinking the Development of Large Language Models from the Causal Perspective: A Legal Text Prediction Case Study | Haotian Chen; Lingwei Zhang; Yiran Liu; Yang Yu   |
| 9912  | Risk-Aware Continuous Control with Neural Contextual Bandits  | Jose A. Ayala-Romero; Andres Garcia-Saavedra; Xavier Costa-Perez                              |
| 9975  | Assume-Guarantee Reinforcement Learning   | Milad Kazemi; Mateo Perez; Fabio Somenzi; Sadegh Soudjani; Ashutosh Trivedi; Alvaro Velasquez |
| 10183 | Coevolutionary Algorithm for Building Robust Decision Trees under Minimax Regret                                    | Adam Żychowski; Andrew Perrault; Jacek Mańdziuk   |
| 10629 | Galileo: General Linear Relaxation Framework for Tightening Robustness Certification of Transformers                | Yun-Ruo Zhang; Lujia Shen; Shanqing GUO; Shouling Ji  |
| 10651 | Divide-and-Aggregate Learning for Evaluating Performance on Unlabeled Data  | Shuyu Miao; Jian Liu; Lin Zheng; Hong Jin   |
| 10738 | Robust Uncertainty Quantification Using Conformalised Monte Carlo Prediction  | Daniel Bethell; Simos Gerasimou; Radu Calinescu   |
| 10756 | Closing the Gap: Achieving Better Accuracy-Robustness Tradeoffs against Query-Based Attacks                         | Pascal Zimmer; Sébastien Andreina; Giorgia Azzurra Marson; Ghassan Karame                     |
| 10916 | Enhancing Off-Policy Constrained Reinforcement Learning through Adaptive Ensemble C Estimation                      | Hengrui Zhang; Youfang Lin; Shuo Shen; Sheng Han; Kai Lv                                      |
| 10939 | Robust Stochastic Graph Generator for Counterfactual Explanations   | Mario Prado Romero; Bardh Prenkaj; Giovanni Stilo   |
| 11084 | Adversarial Initialization with Universal Adversarial Perturbation: A New Approach to Fast Adversarial Training     | Chao Pan; Qing Li; Xin Yao  |



|       |   |  |
|-------|---|--|
| 11313 | Safe Reinforcement Learning with Instantaneous Constraints: The Role of Aggressive                              | Honghao Wei; Xin Liu; Lei Ying   |
| 11483 | Safeguarded Progress in Reinforcement Learning: Safe Bayesian Exploration for Control Policy Synthesis          | Rohan Mitta; Hosein Hasanbeig; Jun Wang; Daniel Kroening; Yiannis Kantaros; Alessandro Abate   |
| 11524 | Solving Non-rectangular Reward-Robust MDPs via Frequency Regularization   | Uri Gadot; Esther Derman; Navdeep Kumar; Maxence Mohamed Elfatih; Kfir Levy; Shie Mannor   |
| 11580 | A Simple and Yet Fairly Effective Defense for Graph Neural Networks   | Sofiane ENNADIR; Yassine ABBAHADDOU; Johannes Lutzeyer; Michalis Vazirgiannis; Henrik Bostrom  |
| 11751 | Hypothesis Testing for Class-Conditional Noise Using Local Maximum Likelihood                                   | Weisong Yang; Rafael Poyiadzi; Niall Twomey; Raul Santos Rodriguez   |
| 11922 | Neural Closure Certificates   | Alireza Nadali; Vishnu Murali; Ashutosh Trivedi; Majid Zamani  |
| 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  |
| 12016 | Stability Analysis of Switched Linear Systems with Neural Lyapunov Functions                                    | Virginie Debauche; Alec Edwards; Raphaël Jungers; Alessandro Abate   |
| 12040 | AdvST: Revisiting Data Augmentations for Single Domain Generalization   | Guangtao Zheng; Mengdi Huai; Aidong Zhang  |
| 12125 | Omega-Regular Decision Processes  | Ernst Moritz Hahn; Mateo Perez; Sven Schewe; Fabio Somenzi; Ashutosh Trivedi; Dominik Wojtczak   |
| 12329 | SentinelLMs: Encrypted Input Adaptation and Fine-Tuning of Language Models for Private and Secure Inference     | Abhijit Mishra; Mingda Li; Soham Deo   |
| 12398 | Providing Fair Recourse over Plausible Groups   | Jayanth Yetukuri; Ian Hardy; Yevgeniy Vorobeychik; Berk Ustun; Yang Liu  |
| 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   |
| 12790 | Chasing Fairness in Graphs: A GNN Architecture Perspective  | Zhimeng Jiang; Xiaotian Han; Chao Fan; Zirui Liu; Na Zou; Ali Mostafavi; Xia Hu  |
| 12862 | Find the Lady: Permutation and Re-synchronization of Deep Neural Networks                                       | Carl De Sousa Trias; Mihai Petru Mitrea; Attilio Fiandrotti; Marco Cagnazzo; Sumanta Chaudhuri; Enzo Tartaglione                         |
| 13175 | The Evidence Contraction Issue in Deep Evidential Regression: Discussion and Solution                           | Yuefei Wu; Bin Shi; Bo Dong; Qinghua Zheng; Hua Wei  |
| 13195 | Robustness Verification of Multi-Class Tree Ensembles   | Laurens Devos; Lorenzo Cascioli; Jesse Davis   |
| 13267 | Beyond Traditional Threats: A Persistent Backdoor Attack on Federated Learning                                  | Tao Liu; Yuhang Zhang; Zhu Feng; Zhiqin Yang; Chen Xu; Dapeng Man; Yang Wu   |
| 13408 | Robust Active Measuring under Model Uncertainty   | Merlijn Krales; Thiago Simão; Jana Tumova; Nils Jansen   |
| 13567 | OUTFOX: LLM-Generated Essay Detection Through In-Context Learning with Adversarially Generated Examples         | Ryuto Koike; Masahiro Kaneko; Naoaki Okazaki   |
| 13647 | All but One: Surgical Erasing with Model Preservation in Text-to-Image Diffusion Models                         | SeungHoo Hong; Juhun Lee; Simon S. Woo   |
| 13716 | Human-Guided Moral Decision Making in Text-Based Games  | Zijing Shi; Meng Fang; Ling Chen; Yali Du; Jun WANG  |
| 13752 | DataElixir: Purifying Poisoned Dataset to Mitigate Backdoor Attacks via Diffusion Models                        | Jiachen Zhou; Peizhuo Lv; Yibing Lan; Guozhu Meng; Kai Chen; Hualong Ma  |
| 13760 | Layer Attack Unlearning: Fast and Accurate Machine Unlearning via Layer Level Attack and Knowledge Distillation | Hyunjune Kim; Sangyong Lee; Simon S. Woo   |