

38th AAAI Conference on Artificial Intelligence (AAAI-24), 36th Conference on Innovative Applications of Artificial Intelligence (IAAI-24), 14th Symposium on Educational Advances in Artificial Intelligence (EAAI-24)

Volume 19: AAAI Special Track

- Safe, Robust and Responsible AI Track

Vancouver, Canada
20 – 27 February 2024

Part 1 of 2

ISBN: 979-8-3313-0102-6

Printed from e-media with permission by:

Curran Associates, Inc.
57 Morehouse Lane
Red Hook, NY 12571



Some format issues inherent in the e-media version may also appear in this print version.

Copyright© (2024) by Association for the Advancement of Artificial Intelligence
All rights reserved.

Printed with permission by Curran Associates, Inc. (2024)

For permission requests, please contact Association for the Advancement of Artificial Intelligence
at the address below.

Association for the Advancement of Artificial Intelligence
2275 East Bayshore Road
Suite 160
Palo Alto, California 94303
USA

Phone: 1-650-328-3123
Fax: 1-650-321-4457

<https://aaai.org/Press/press.php>

Additional copies of this publication are available from:

Curran Associates, Inc.
57 Morehouse Lane
Red Hook, NY 12571 USA
Phone: 845-758-0400
Fax: 845-758-2633
Email: curran@proceedings.com
Web: www.proceedings.com

TABLE OF CONTENTS

PART 1

AAAI TECHNICAL TRACK ON SAFE, ROBUST AND RESPONSIBLE AI TRACK

ImageCaptioner2: Image Captioner for Image Captioning Bias Amplification Assessment.....	20902
<i>Eslam Abdelrahman, Pengzhan Sun, Li Erran Li, Mohamed Elhoseiny</i>	
A Framework for Data-Driven Explainability in Mathematical Optimization.....	20912
<i>Kevin-Martin Aigner, Marc Goerigk, Michael Hartisch, Frauke Liers, Arthur Miehlich</i>	
On the Importance of Application-Grounded Experimental Design for Evaluating Explainable ML Methods.....	20921
<i>Kasun Amarasinghe, Kit T. Rodolfa, Sérgio Jesus, Valerie Chen, Vladimir Balayan, Pedro Saleiro, Pedro Bizarro, Ameet Talwalkar, Rayid Ghani</i>	
Risk-Aware Continuous Control with Neural Contextual Bandits.....	20930
<i>Jose A. Ayala-Romero, Andres Garcia-Saavedra, Xavier Costa-Perez</i>	
Robust Uncertainty Quantification Using Conformalised Monte Carlo Prediction.....	20939
<i>Daniel Bethell, Simos Gerasimou, Radu Calinescu</i>	
CCTR: Calibrating Trajectory Prediction for Uncertainty-Aware Motion Planning in Autonomous Driving.....	20949
<i>Chengtai Cao, Xinhong Chen, Jianping Wang, Qun Song, Rui Tan, Yung-Hui Li</i>	
Rethinking the Development of Large Language Models from the Causal Perspective: A Legal Text Prediction Case Study.....	20958
<i>Haotian Chen, Lingwei Zhang, Yiran Liu, Yang Yu</i>	
Truth Forest: Toward Multi-Scale Truthfulness in Large Language Models Through Intervention Without Tuning.....	20967
<i>Zhongzhi Chen, Xingwu Sun, Xianfeng Jiao, Fengzong Lian, Zhanhui Kang, Di Wang, Chengzhong Xu</i>	
Constrained Meta-Reinforcement Learning for Adaptable Safety Guarantee with Differentiable Convex Programming.....	20975
<i>Minjae Cho, Chuangchuang Sun</i>	
Conformal Prediction Regions for Time Series Using Linear Complementarity Programming.....	20984
<i>Matthew Cleaveland, Insup Lee, George J. Pappas, Lars Lindemann</i>	
TTTS: Tree Test Time Simulation for Enhancing Decision Tree Robustness Against Adversarial Examples.....	20993
<i>Seffi Cohen, Ofir Arbili, Yisroel Mirsky, Lior Rokach</i>	
Find the Lady: Permutation and Re-Synchronization of Deep Neural Networks.....	21001
<i>Carl De Sousa Trias, Mihai Petru Mitrea, Attilio Fiandrotti, Marco Cagnazzo, Sumanta Chaudhuri, Enzo Tartaglione</i>	
Stability Analysis of Switched Linear Systems with Neural Lyapunov Functions.....	21010
<i>Virginie Debauche, Alec Edwards, Raphaël M. Jungers, Alessandro Abate</i>	

Robustness Verification of Multi-Class Tree Ensembles.....	21019
<i>Laurens Devos, Lorenzo Cascioli, Jesse Davis</i>	
P2BPO: Permeable Penalty Barrier-Based Policy Optimization for Safe RL	21029
<i>Sumanta Dey, Pallab Dasgupta, Soumyajit Dey</i>	
Trade-Offs in Fine-Tuned Diffusion Models Between Accuracy and Interpretability.....	21037
<i>Mischa Dombrowski, Hadrien Reynaud, Johanna P. Müller, Matthew Baugh, Bernhard Kainz</i>	
From Hope to Safety: Unlearning Biases of Deep Models Via Gradient Penalization in Latent Space	21046
<i>Maximilian Dreyer, Frederik Pahde, Christopher J. Anders, Wojciech Samek, Sebastian Lapuschkin</i>	
Automatically Testing Functional Properties of Code Translation Models	21055
<i>Hasan Ferit Eniser, Valentin Wüstholtz, Maria Christakis</i>	
A Simple and Yet Fairly Effective Defense for Graph Neural Networks	21063
<i>Sofiane Ennadir, Yassine Abbahaddou, Johannes F. Lutzeyer, Michalis Vazirgiannis, Henrik Boström</i>	
Invisible Backdoor Attack Against 3D Point Cloud Classifier in Graph Spectral Domain	21072
<i>Linkun Fan, Fazhi He, Tongzhen Si, Wei Tang, Bing Li</i>	
CASE: Exploiting Intra-Class Compactness and Inter-Class Separability of Feature Embeddings for Out-Of-Distribution Detection	21081
<i>Shuai Feng, Pengsheng Jin, Chongjun Wang</i>	
Solving Non-Rectangular Reward-Robust MDPs Via Frequency Regularization.....	21090
<i>Uri Gadot, Esther Derman, Navdeep Kumar, Maxence Mohamed Elfatih, Kfir Levy, Shie Mannor</i>	
Balance Reward and Safety Optimization for Safe Reinforcement Learning: A Perspective of Gradient Manipulation.....	21099
<i>Shangding Gu, Bilgehan Sel, Yuhao Ding, Lu Wang, Qingwei Lin, Ming Jin, Alois Knoll</i>	
π -Light: Programmatic Interpretable Reinforcement Learning for Resource-Limited Traffic Signal Control.....	21107
<i>Yin Gu, Kai Zhang, Qi Liu, Weibo Gao, Longfei Li, Jun Zhou</i>	
Generative Model for Decision Trees.....	21116
<i>Riccardo Guidotti, Anna Monreale, Mattia Setzu, Giulia Volpi</i>	
Omega-Regular Decision Processes.....	21125
<i>Ernst Moritz Hahn, Mateo Perez, Sven Schewe, Fabio Somenzi, Ashutosh Trivedi, Dominik Wojtczak</i>	
Provable Robustness Against a Union of L_0 Adversarial Attacks.....	21134
<i>Zayd Hammoudeh, Daniel Lowd</i>	
All but One: Surgical Concept Erasing with Model Preservation in Text-To-Image Diffusion Models.....	21143
<i>SeungHoo Hong, Juhun Lee, Simon S. Woo</i>	
Towards Efficient Verification of Quantized Neural Networks.....	21152
<i>Pei Huang, Haoze Wu, Yuting Yang, Ieva Daukantas, Min Wu, Yedi Zhang, Clark Barrett</i>	

On the Concept Trustworthiness in Concept Bottleneck Models	21161
<i>Qihan Huang, Jie Song, Jingwen Hu, Haofei Zhang, Yong Wang, Mingli Song</i>	
Personalization as a Shortcut for Few-Shot Backdoor Attack Against Text-To-Image Diffusion Models.....	21169
<i>Yihao Huang, Felix Juefei-Xu, Qing Guo, Jie Zhang, Yutong Wu, Ming Hu, Tianlin Li, Geguang Pu, Yang Liu</i>	
Stronger and Transferable Node Injection Attacks	21179
<i>Samyak Jain, Tanima Dutta</i>	
Learning Fair Policies for Multi-Stage Selection Problems from Observational Data.....	21188
<i>Zhuangzhuang Jia, Grani A. Hanasusanto, Phebe Vayanos, Weijun Xie</i>	
NeRFail: Neural Radiance Fields-Based Multiview Adversarial Attack.....	21197
<i>Wenxiang Jiang, Hanwei Zhang, Xi Wang, Zhongwen Guo, Hao Wang</i>	
Analysis of Differentially Private Synthetic Data: A Measurement Error Approach.....	21206
<i>Yangdi Jiang, Yi Liu, Xiaodong Yan, Anne-Sophie Charest, Linglong Kong, Bei Jiang</i>	
Chasing Fairness in Graphs: A GNN Architecture Perspective	21214
<i>Zhimeng Jiang, Xiaotian Han, Chao Fan, Zirui Liu, Na Zou, Ali Mostafavi, Xia Hu</i>	
Assume-Guarantee Reinforcement Learning.....	21223
<i>Milad Kazemi, Mateo Perez, Fabio Somenzi, Sadegh Soudjani, Ashutosh Trivedi, Alvaro Velasquez</i>	
DeepBern-Nets: Taming the Complexity of Certifying Neural Networks Using Bernstein Polynomial Activations and Precise Bound Propagation.....	21232
<i>Haitham Khedr, Yasser Shoukry</i>	
Layer Attack Unlearning: Fast and Accurate Machine Unlearning Via Layer Level Attack and Knowledge Distillation.....	21241
<i>Hyunjune Kim, Sangyong Lee, Simon S. Woo</i>	
Quilt: Robust Data Segment Selection Against Concept Drifts.....	21249
<i>Minsu Kim, Seong-Hyeon Hwang, Steven Euijong Whang</i>	
OUTFOX: LLM-Generated Essay Detection Through In-Context Learning with Adversarially Generated Examples.....	21258
<i>Ryuto Koike, Masahiro Kaneko, Naoaki Okazaki</i>	
Accelerating Adversarially Robust Model Selection for Deep Neural Networks Via Racing.....	21267
<i>Matthias König, Holger H. Hoos, Jan N. van Rijn</i>	
Robust Active Measuring Under Model Uncertainty	21276
<i>Merlijn Krale, Thiago D. Simão, Jana Tumova, Nils Jansen</i>	
Towards Large Certified Radius in Randomized Smoothing Using Quasiconcave Optimization.....	21285
<i>Bo-Han Kung, Shang-Tse Chen</i>	
Contrastive Credibility Propagation for Reliable Semi-Supervised Learning.....	21294
<i>Brody Kutt, Pralay Ramteke, Xavier Mignot, Pamela Toman, Nandini Ramanan, Sujit Rokka Chhetri, Shan Huang, Min Du, William Hewlett</i>	

Exponent Relaxation of Polynomial Zonotopes and Its Applications in Formal Neural Network Verification	21304
<i>Tobias Ladner, Matthias Althoff</i>	
I Prefer Not to Say: Protecting User Consent in Models with Optional Personal Data	21312
<i>Tobias Leemann, Martin Pawelczyk, Christian Thomas Eberle, Gjergji Kasneci</i>	
Promoting Counterfactual Robustness Through Diversity	21322
<i>Francesco Leofante, Nico Potyka</i>	
Revisiting the Information Capacity of Neural Network Watermarks: Upper Bound Estimation and Beyond	21331
<i>Fangqi Li, Haodong Zhao, Wei Du, Shilin Wang</i>	
PointCVaR: Risk-Optimized Outlier Removal for Robust 3D Point Cloud Classification	21340
<i>Xinke Li, Junchi Lu, Henghui Ding, Changsheng Sun, Joey Tianyi Zhou, Yeow Meng Chee</i>	
Game-Theoretic Unlearnable Example Generator	21349
<i>Shuang Liu, Yihan Wang, Xiao-Shan Gao</i>	
Beyond Traditional Threats: A Persistent Backdoor Attack on Federated Learning.....	21359
<i>Tao Liu, Yuhang Zhang, Zhu Feng, Zhiqin Yang, Chen Xu, Dapeng Man, Wu Yang</i>	
Handling Long and Richly Constrained Tasks Through Constrained Hierarchical Reinforcement Learning	21368
<i>Yuxiao Lu, Arunesh Sinha, Pradeep Varakantham</i>	
Combining Graph Transformers Based Multi-Label Active Learning and Informative Data Augmentation for Chest Xray Classification.....	21378
<i>Dwarikanath Mahapatra, Behzad Bozorgtabar, Zongyuan Ge, Mauricio Reyes, Jean-Philippe Thiran</i>	

PART 2

Enumerating Safe Regions in Deep Neural Networks with Provable Probabilistic Guarantees.....	21387
<i>Luca Marzari, Davide Corsi, Enrico Marchesini, Alessandro Farinelli, Ferdinando Cicalese</i>	
Divide-And-Aggregate Learning for Evaluating Performance on Unlabeled Data.....	21395
<i>Shuyu Miao, Jian Liu, Lin Zheng, Hong Jin</i>	
SentinelLMs: Encrypted Input Adaptation and Fine-Tuning of Language Models for Private and Secure Inference	21403
<i>Abhijit Mishra, Mingda Li, Soham Deo</i>	
Safeguarded Progress in Reinforcement Learning: Safe Bayesian Exploration for Control Policy Synthesis.....	21412
<i>Rohan Mitta, Hosein Hasanbeig, Jun Wang, Daniel Kroening, Yiannis Kantaros, Alessandro Abate</i>	
Feature Unlearning for Pre-Trained GANs and VAEs.....	21420
<i>Saemi Moon, Seunghyuk Cho, Dongwoo Kim</i>	
Reward Certification for Policy Smoothed Reinforcement Learning.....	21429
<i>Ronghui Mu, Leandro Soriano Marcolino, Yanghao Zhang, Tianle Zhang, Xiaowei Huang, Wenjie Ruan</i>	

EncryIP: A Practical Encryption-Based Framework for Model Intellectual Property Protection.....	21438
<i>Xin Mu, Yu Wang, Zhengan Huang, Junzuo Lai, Yehong Zhang, Hui Wang, Yue Yu</i>	
Neural Closure Certificates	21446
<i>Alireza Nadali, Vishnu Murali, Ashutosh Trivedi, Majid Zamani</i>	
SocialStigmaQA: A Benchmark to Uncover Stigma Amplification in Generative Language Models	21454
<i>Manish Nagireddy, Lamogha Chiazor, Moninder Singh, Ioana Baldini</i>	
MaxEnt Loss: Constrained Maximum Entropy for Calibration Under Out-Of-Distribution Shift	21463
<i>Dexter Neo, Stefan Winkler, Tsuhan Chen</i>	
ORES: Open-Vocabulary Responsible Visual Synthesis	21473
<i>Minheng Ni, Chenfei Wu, Xiaodong Wang, Shengming Yin, Lijuan Wang, Zicheng Liu, Nan Duan</i>	
Q-SENN: Quantized Self-Explaining Neural Networks.....	21482
<i>Thomas Norrenbrock, Marco Rudolph, Bodo Rosenhahn</i>	
Understanding Likelihood of Normalizing Flow and Image Complexity Through the Lens of Out-Of-Distribution Detection.....	21492
<i>Genki Osada, Tsubasa Takahashi, Takashi Nishide</i>	
Adversarial Initialization with Universal Adversarial Perturbation: A New Approach to Fast Adversarial Training	21501
<i>Chao Pan, Qing Li, Xin Yao</i>	
A PAC Learning Algorithm for LTL and Omega-Regular Objectives in MDPs.....	21510
<i>Mateo Perez, Fabio Somenzi, Ashutosh Trivedi</i>	
Robust Stochastic Graph Generator for Counterfactual Explanations.....	21518
<i>Mario Alfonso Prado-Romero, Bardh Prenkaj, Giovanni Stilo</i>	
Visual Adversarial Examples Jailbreak Aligned Large Language Models	21527
<i>Xiangyu Qi, Kaixuan Huang, Ashwinee Panda, Peter Henderson, Mengdi Wang, Prateek Mittal</i>	
Dissenting Explanations: Leveraging Disagreement to Reduce Model Overreliance	21537
<i>Omer Reingold, Judy Hanwen Shen, Aditi Talati</i>	
I-CEE: Tailoring Explanations of Image Classification Models to User Expertise	21545
<i>Yao Rong, Peizhu Qian, Vaibhav Unhelkar, Enkelejda Kasneci</i>	
A Simple and Practical Method for Reducing the Disparate Impact of Differential Privacy	21554
<i>Lucas Rosenblatt, Julia Stoyanovich, Christopher Musco</i>	
Interpretability Benchmark for Evaluating Spatial Misalignment of Prototypical Parts Explanations.....	21563
<i>Mikolaj Sacha, Bartosz Jura, Dawid Rymarczyk, Lukasz Struski, Jacek Tabor, Bartosz Zieliński</i>	
Human-Guided Moral Decision Making in Text-Based Games	21574
<i>Zijing Shi, Meng Fang, Ling Chen, Yali Du, Jun Wang</i>	
Towards Fairer Centroids in K-Means Clustering	21583
<i>Stanley Simoes, Deepak P, Muiris MacCarthaigh</i>	

Toward Robustness in Multi-Label Classification: A Data Augmentation Strategy Against Imbalance and Noise	21592
<i>Hwanjun Song, Minseok Kim, Jae-Gil Lee</i>	
Bidirectional Contrastive Split Learning for Visual Question Answering	21602
<i>Yuwei Sun, Hideya Ochiai</i>	
Quantile-Based Maximum Likelihood Training for Outlier Detection	21610
<i>Masoud Taghikhah, Nishant Kumar, Siniša Šegvić, Abouzar Eslami, Stefan Gumhold</i>	
Sparsity-Guided Holistic Explanation for LLMs with Interpretable Inference-Time Intervention	21619
<i>Zhen Tan, Tianlong Chen, Zhenyu Zhang, Huan Liu</i>	
Toward More Generalized Malicious URL Detection Models	21628
<i>Yun-Da Tsai, Cayon Liow, Yin Sheng Siang, Shou-De Lin</i>	
Self-Supervised Likelihood Estimation with Energy Guidance for Anomaly Segmentation in Urban Scenes.....	21637
<i>Yuanpeng Tu, Yuxi Li, Boshen Zhang, Liang Liu, Jiangning Zhang, Yabiao Wang, Cairong Zhao</i>	
Pure-Past Action Masking	21646
<i>Giovanni Varricchione, Natasha Alechina, Mehdi Dastani, Giuseppe De Giacomo, Brian Logan, Giuseppe Perelli</i>	
Long-Term Safe Reinforcement Learning with Binary Feedback.....	21656
<i>Akifumi Wachi, Wataru Hashimoto, Kazumune Hashimoto</i>	
Identifying Reasons for Bias: An Argumentation-Based Approach	21664
<i>Madeleine Waller, Odinaldo Rodrigues, Oana Cocarascu</i>	
Would You Like Your Data to Be Trained? a User Controllable Recommendation Framework	21673
<i>Lei Wang, Xu Chen, Zhenhua Dong, Quanyu Dai</i>	
Moderate Message Passing Improves Calibration: A Universal Way to Mitigate Confidence Bias in Graph Neural Networks.....	21681
<i>Min Wang, Hao Yang, Jincui Huang, Qing Cheng</i>	
Generating Diagnostic and Actionable Explanations for Fair Graph Neural Networks	21690
<i>Zhenzhong Wang, Qingyuan Zeng, Wanyu Lin, Min Jiang, Kay Chen Tan</i>	
Physics-Informed Representation and Learning: Control and Risk Quantification.....	21699
<i>Zhuoyuan Wang, Reece Keller, Xiyu Deng, Kenta Hoshino, Takashi Tanaka, Yorie Nakahira</i>	
Safe Reinforcement Learning with Instantaneous Constraints: The Role of Aggressive Exploration.....	21708
<i>Honghao Wei, Xin Liu, Lei Ying</i>	
Concealing Sensitive Samples Against Gradient Leakage in Federated Learning	21717
<i>Jing Wu, Munawar Hayat, Mingyi Zhou, Mehrtash Harandi</i>	
The Evidence Contraction Issue in Deep Evidential Regression: Discussion and Solution	21726
<i>Yuefei Wu, Bin Shi, Bo Dong, Qinghua Zheng, Hua Wei</i>	
Byzantine-Robust Decentralized Learning Via Remove-Then-Clip Aggregation	21735
<i>Caiyi Yang, Javad Ghaderi</i>	

Hypothesis Testing for Class-Conditional Noise Using Local Maximum Likelihood	21744
<i>Weisong Yang, Rafael Poyiadzi, Niall Twomey, Raul Santos-Rodriguez</i>	
Providing Fair Recourse Over Plausible Groups.....	21753
<i>Jayanth Yetukuri, Ian Hardy, Yevgeniy Vorobeychik, Berk Ustun, Yang Liu</i>	
Representation-Based Robustness in Goal-Conditioned Reinforcement Learning	21761
<i>Xiangyu Yin, Sihao Wu, Jiaxu Liu, Meng Fang, Xingyu Zhao, Xiaowei Huang, Wenjie Ruan</i>	
Enhancing Off-Policy Constrained Reinforcement Learning Through Adaptive Ensemble C Estimation.....	21770
<i>Hengrui Zhang, Youfang Lin, Shuo Shen, Sheng Han, Kai Lv</i>	
Efficient Toxic Content Detection by Bootstrapping and Distilling Large Language Models	21779
<i>Jiang Zhang, Qiong Wu, Yiming Xu, Cheng Cao, Zheng Du, Konstantinos Psounis</i>	
LR-XFL: Logical Reasoning-Based Explainable Federated Learning	21788
<i>Yanci Zhang, Han Yu</i>	
GaLileo: General Linear Relaxation Framework for Tightening Robustness Certification of Transformers.....	21797
<i>Yunruo Zhang, Lujia Shen, Shanqing Guo, Shouling Ji</i>	
A Huber Loss Minimization Approach to Byzantine Robust Federated Learning	21806
<i>Puning Zhao, Fei Yu, Zhiguo Wan</i>	
Responsible Bandit Learning Via Privacy-Protected Mean-Volatility Utility	21815
<i>Shanshan Zhao, Wenhai Cui, Bei Jiang, Linglong Kong, Xiaodong Yan</i>	
UMA: Facilitating Backdoor Scanning Via Unlearning-Based Model Ablation.....	21823
<i>Yue Zhao, Congyi Li, Kai Chen</i>	
AdvST: Revisiting Data Augmentations for Single Domain Generalization.....	21832
<i>Guangtao Zheng, Mengdi Huai, Aidong Zhang</i>	
Can LLM Replace Stack Overflow? a Study on Robustness and Reliability of Large Language Model Code Generation	21841
<i>Li Zhong, Zilong Wang</i>	
DataElixir: Purifying Poisoned Dataset to Mitigate Backdoor Attacks Via Diffusion Models	21850
<i>Jiachen Zhou, Peizhuo Lv, Yibing Lan, Guozhu Meng, Kai Chen, Hualong Ma</i>	
Closing the Gap: Achieving Better Accuracy-Robustness Tradeoffs Against Query-Based Attacks	21859
<i>Pascal Zimmer, Sébastien Andreina, Giorgia Azzurra Marson, Ghassan Karame</i>	
Coevolutionary Algorithm for Building Robust Decision Trees Under Minimax Regret	21869
<i>Adam Żychowski, Andrew Perrault, Jacek Mańdziuk</i>	

Author Index