

Table of Contents

Session 1: Adversarial Machine Learning	
Bio-Inspired Adversarial Attack Against Deep Neural Networks	1
<i>Bowei Xi, Yujie Chen, Fei Fan, Zhan Tu and Xinyan Deng</i>	
Adversarial Image Translation: Unrestricted Adversarial Examples in Face Recognition Systems	6
<i>Kazuya Kakizaki and Kosuke Yoshida</i>	
<hr/>	
Session 2: Assurance Cases for AI-based Systems	
Hazard Contribution Modes of Machine Learning Components	14
<i>Ewen Denney, Ganesh Pai and Colin Smith</i>	
Assurance Argument Patterns and Processes for Machine Learning in Safety-Related Systems	23
<i>Chiara Picardi, Colin Paterson, Richard Hawkins, Radu Calinescu and Ibrahim Habli</i>	
<hr/>	
Session 3: Considerations for the AI Safety Landscape	
Founding The Domain of AI Forensics	31
<i>Vahid Behzadan and Ibrahim Baggili</i>	
Exploring AI Safety in Degrees: Generality, Capability and Control	36
<i>John Burden and José Hernández-Orallo</i>	
<hr/>	
Session 4: Fairness and Bias	
Fair Enough: Improving Fairness in Budget-Constrained Decision Making Using Confidence Thresholds	41
<i>Michiel Bakker, Humberto Riveron Valdes, Duy Patrick Tu, Krishna Gummadi, Kush Varshney, Adrian Weller and Alex Pentland</i>	
A Study on Multimodal and Interactive Explanations for Visual Question Answering	54
<i>Kamran Alipour, Jurgen P. Schulze, Yi Yao, Avi Ziskind and Giedrius Burachas</i>	
You Shouldn't Trust Me: Learning Models Which Conceal Unfairness From Multiple Explanation Methods	63
<i>Botty Dimanov, Umang Bhatt, Mateja Jamnik and Adrian Weller</i>	
<hr/>	
Session 5: Uncertainty and Safe AI	
A High Probability Safety Guarantee with Shifted Neural Network Surrogates	74
<i>Mélanie Ducoffe, Jayant Sen Gupta and Sebastien Gerchinovitz</i>	
Benchmarking Uncertainty Estimation Methods for Deep Learning With Safety-Related Metrics	83
<i>Maximilian Henne, Adrian Schwaiger, Karsten Roscher and Gereon Weiss</i>	

PURSS: Towards Perceptual Uncertainty Aware Responsibility Sensitive Safety with ML . . .	91
<i>Rick Salay, Krzysztof Czarnecki, Maria Elli, Ignacio Alvarez, Sean Sedwards and Jack Weast</i>	

Poster Papers

Simple Continual Learning Strategies for Safer Classifiers	96
<i>Ashish Gaurav, Sachin Vernekar, Jaeyoung Lee, Vahdat Abdelzad, Krzysztof Czarnecki and Sean Sedwards</i>	
Fair Representation for Safe Artificial Intelligence via Adversarial Learning of Unbiased Information Bottleneck	105
<i>Jin-Young Kim and Sung-Bae Cho</i>	
Out-of-Distribution Detection with Likelihoods Assigned by Deep Generative Models Using Multimodal Prior Distributions	113
<i>Ryo Kamoi and Kei Kobayashi</i>	
SafeLife 1.0: Exploring Side Effects in Complex Environments	117
<i>Carroll Wainwright and Peter Eckersley</i>	
(When) Is Truth-telling Favored in AI Debate?	128
<i>Vojtech Kovarik and Ryan Carey</i>	
NewsBag: A Benchmark Multimodal Dataset for Fake News Detection	138
<i>Sarthak Jindal, Raghav Sood, Richa Singh, Mayank Vatsa and Tanmoy Chakraborty</i>	
Algorithmic Discrimination: Formulation and Exploration in Deep Learning-based Face Biometrics	146
<i>Ignacio Serna, Aythami Morales, Julian Fierrez, Manuel Cebrian, Nick Obradovich and Iyad Rahwan</i>	
Guiding Safe Reinforcement Learning Policies Using Structured Language Constraints . . .	153
<i>Bharat Prakash, Nicholas Waytowich, Ashwinkumar Ganesan, Tim Oates and Tinoosh Mohsenin</i>	
Practical Solutions for Machine Learning Safety in Autonomous Vehicles	162
<i>Sina Mohseni, Mandar Pitale, Vasu Singh and Zhangyang Wang</i>	
Continuous Safe Learning Based on First Principles and Constraints for Autonomous Driving	170
<i>Lifeng Liu, Yingxuan Zhu, Tim Yuan and Jian Li</i>	
Recurrent Neural Network Properties and their Verification with Monte Carlo Techniques	178
<i>Dmitry Vengertsev and Elena Sherman</i>	
Toward Operational Safety Verification Via Hybrid Automata Mining Using I/O Traces of AI-Enabled CPS	186
<i>Imane Lamrani, Ayan Banerjee and Sandeep Gupta</i>	