## Table of Contents

### Session 1: Adversarial Machine Learning

- Bio-Inspired Adversarial Attack Against Deep Neural Networks .............................................. 1
  *Bowei Xi, Yujie Chen, Fei Fan, Zhan Tu and Xinyan Deng*
- Adversarial Image Translation: Unrestricted Adversarial Examples in Face Recognition Systems .................................................. 6
  *Kazuya Kakizaki and Kosuke Yoshida*

### Session 2: Assurance Cases for AI-based Systems

- Hazard Contribution Modes of Machine Learning Components ................................................. 14
  *Ewen Denney, Ganesh Pai and Colin Smith*
- Assurance Argument Patterns and Processes for Machine Learning in Safety-Related Systems .................................................................................. 23
  *Chiara Picardi, Colin Paterson, Richard Hawkins, Radu Calinescu and Ibrahim Habli*

### Session 3: Considerations for the AI Safety Landscape

- Founding The Domain of AI Forensics ......................................................................................... 31
  *Vahid Behzadan and Ibrahim Baggili*
- Exploring AI Safety in Degrees: Generality, Capability and Control ........................................ 36
  *John Burden and José Hernández-Orallo*

### Session 4: Fairness and Bias

- Fair Enough: Improving Fairness in Budget-Constrained Decision Making Using Confidence Thresholds ................................................................. 41
  *Michiel Bakker, Humberto Riveron Valdes, Duy Patrick Tu, Krishna Gummad, Kush Varshney, Adrian Weller and Alex Pentland*
- A Study on Multimodal and Interactive Explanations for Visual Question Answering ..... 54
  *Kamran Alipour, Jurgen P. Schulze, Yi Yao, Avi Ziskind and Giedrius Burachas*
- You Shouldn’t Trust Me: Learning Models Which Conceal Unfairness From Multiple Explanation Methods ................................................................. 63
  *Botty Dimanov, Umang Bhatt, Mateja Jamnik and Adrian Weller*

### Session 5: Uncertainty and Safe AI

- A High Probability Safety Guarantee with Shifted Neural Network Surrogates ..................... 74
  *Mélanie Ducoffe, Jayant Sen Gupta and Sebastien Gerchinovitz*
- Benchmarking Uncertainty Estimation Methods for Deep Learning With Safety-Related Metrics .................................................................................. 83
  *Maximilian Henne, Adrian Schwaiger, Karsten Roscher and Gereon Weiss*
# SafeAI 2020 Table of Contents

**PURSS:** Towards Perceptual Uncertainty Aware Responsibility Sensitive Safety with ML  
*Rick Salay, Krzysztof Czarnecki, Maria Elli, Igancio Alvarez, Sean Sedwards and Jack Weast*

## Poster Papers

**Simple Continual Learning Strategies for Safer Classifiers**  
*Ashish Gaurav, Sachin Vernekar, Jaeyoung Lee, Vahdat Abdelzad, Krzysztof Czarnecki and Sean Sedwards*

**Fair Representation for Safe Artificial Intelligence via Adversarial Learning of Unbiased Information Bottleneck**  
*Jin-Young Kim and Sung-Bae Cho*

**Out-of-Distribution Detection with Likelihoods Assigned by Deep Generative Models Using Multimodal Prior Distributions**  
*Ryo Kamoi and Kei Kobayashi*

**SafeLife 1.0: Exploring Side Effects in Complex Environments**  
*Carroll Wainwright and Peter Eckersley*

**(When) Is Truth-telling Favored in AI Debate?**  
*Vojtech Kovarik and Ryan Carey*

**NewsBag: A Benchmark Multimodal Dataset for Fake News Detection**  
*Sarthak Jindal, Raghav Sood, Richa Singh, Mayank Vatsa and Tanmoy Chakraborty*

**Algorithmic Discrimination: Formulation and Exploration in Deep Learning-based Face Biometrics**  
*Ignacio Serna, Aythami Morales, Julian Fierrez, Manuel Cebrian, Nick Obradovich and Iyad Rahwan*

**Guiding Safe Reinforcement Learning Policies Using Structured Language Constraints**  
*Bharat Prakash, Nicholas Waytowich, Ashwinkumar Ganesan, Tim Oates and Tinoosh Mohsenin*

**Practical Solutions for Machine Learning Safety in Autonomous Vehicles**  
*Sina Mohseni, Mandar Pitale, Vasu Singh and Zhangyang Wang*

**Continuous Safe Learning Based on First Principles and Constraints for Autonomous Driving**  
*Lifeng Liu, Yingzuan Zhu, Tim Yuan and Jian Li*

**Recurrent Neural Network Properties and their Verification with Monte Carlo Techniques**  
*Dmitry Vengertsev and Elena Sherman*

**Toward Operational Safety Verification Via Hybrid Automata Mining Using I/O Traces of AI-Enabled CPS**  
*Imane Lamrani, Ayan Banerjee and Sandeep Gupta*

---

2