SafeAI 2019 – Selected Submissions

FULL PRESENTATION (39% acceptance rate)

- Bryant Chen, Wilka Carvalho, Nathalie Baracaldo, Heiko Ludwig, Benjamin Edwards, Taesung Lee, Ian Molloy and Biplav Srivastava. Detecting Backdoor Attacks on Deep Neural Networks by Activation Clustering
- Alec Banks and Rob Ashmore. Requirements Assurance in Machine Learning
- Shih-Yun Lo, Shani Alkoby and Peter Stone. Robust Motion Planning and Safety Benchmarking in Human Workspaces
- Jose Hernandez-Orallo, Fernando Martínez-Plumed, Shahar Avin and Sean O Heigeartaigh. Surveying Safety-relevant AI Characteristics
- Xin Liu, Huanrui Yang, Ziwei Liu, Linghao Song, Yiran Chen and Hai Li. DPATCH: An Adversarial Patch Attack on Object Detectors
- Katja Auernhammer, Ramin Tavakoli Kolagari and Markus Zoppelt. Attacks on Machine Learning: Lurking Danger for Accountability
- Saasha Nair, Sina Shafaei, Stefan Kugele, Mohd Hafeez Osman and Alois Knoll. Monitoring Safety of Autonomous Vehicles with Crash Prediction Networks
- Sandhya Saisubramanian and Shlomo Zilberstein. Minimizing the Negative Side Effects of Planning with Reduced Models
- Lê Nguyên Hoang. Towards Robust Value Loading
- Philip Koopman and Frank Fratrik. How Many Operational Design Domains, Objects, and Events?
- Gopal Sarma, Adam Safron and Nick Hay. Integrative Biological Simulation, Neuropsychology, and AI Safety
- Mark Riedl and Brent Harrison. Enter the Matrix: Safely Interruptible Autonomous Systems via Virtualization

TALK (+ paper)

- Peter Eckersley. Impossibility and Uncertainty Theorems in AI Value Alignment (or why your AGI should not have a utility function)

POSTERS (+2 minutes pitch and short Poster Papers)

- Frank Rudzicz, P Alison Paprica and Marta Janczarski. Towards international standards for evaluating machine learning
- Vahid Behzadan, Roman V. Yampolskiy and Arslan Munir. Emergence of Addictive Behaviors in Reinforcement Learning Agents
- Lei Cui. Improving Robustness of Neural Networks using Distance Metric Learning
- Philip Amortila, Marc G. Bellemare, Prakash Panangaden and Doina Precup. Temporally Extended Metrics for Markov Decision Processes
- Huanrui Yang, Jingchi Zhang, Hsin-Pai Cheng, Wenhan Wang, Yiran Chen and Hai Li. Bamboo: Ball-Shape Data Augmentation Against Adversarial Attacks from All Directions
- Travis Mandel, Jahnu Best, Randall Tanaka, Hiram Temple, Chansen Haili and Roy Szeto. Exploring Interfaces to Democratize AI Constraint Generation

- Saerom Park, Jaewook Lee, Jung Hee Cheon, Juhee Lee, Jaeyun Kim and Junyoung Byun. Security-preserving Support Vector Machine with Fully Homomorphic Encryption
- Tao Li, Lei Lin and Siyuan Gong. AutoMPC: Efficient Multi-Party Computation for Secure and Privacy-Preserving Cooperative Control of Connected Autonomous Vehicles
- Bence Cserna, William Doyle, Tianyi Gu and Wheeler Ruml. Safe Temporal Planning for Urban Driving
- Kacper Sokol and Peter Flach. Counterfactual Explanations of Machine Learning Predictions: Opportunities and Challenges for AI Safety
- Yi Zeng, Enmeng Lu and Cunqing Huangfu. Linking Artificial Intelligence Principles