

Table of Contents

Invited Talk

Impossibility and Uncertainty Theorems in AI Value Alignment (or why your AGI should not have a utility function).....	1
<i>Peter Eckersley</i>	

Session 1: Safe Planning and Operation of Autonomous Systems

Minimizing the Negative Side Effects of Planning with Reduced Models	9
<i>Sandhya Saisubramanian and Shlomo Zilberstein</i>	
Robust Motion Planning and Safety Benchmarking in Human Workspaces.....	16
<i>Shih-Yun Lo, Shani Alkoby and Peter Stone</i>	
Enter the Matrix: Safely Interruptible Autonomous Systems via Virtualization	25
<i>Mark Riedl and Brent Harrison</i>	

Session 2: New Paradigms in AI and AGI Safety

Towards Robust End-to-End Alignment	31
<i>Lê Nguyễn Hoàng</i>	
Integrative Biological Simulation, Neuropsychology, and AI Safety	40
<i>Gopal Sarma, Adam Safron and Nick Hay</i>	

Session 3: Safety in Automated Driving

How Many Operational Design Domains, Objects, and Events?	45
<i>Philip Koopman and Frank Fratrik</i>	
Monitoring Safety of Autonomous Vehicles with Crash Prediction Networks	49
<i>Saasha Nair, Sina Shafaei, Stefan Kugele, Mohd Hafeez Osman and Alois Knoll</i>	

Session 4: Safety-Related AI Requirements and Characteristics

Requirements Assurance in Machine Learning	53
<i>Alec Banks and Rob Ashmore</i>	
Surveying Safety-relevant AI Characteristics.....	57
<i>Jose Hernandez-Orallo, Fernando Martínez-Plumed, Shahar Avin and Sean O Heigeartaigh</i>	

Session 5: Adversarial Machine Learning

Detecting Backdoor Attacks on Deep Neural Networks by Activation Clustering.....	66
<i>Bryant Chen, Wilka Carvalho, Nathalie Baracaldo, Heiko Ludwig, Benjamin Edwards, Taesung Lee, Ian Molloy and Biplav Srivastava</i>	
DPATCH: An Adversarial Patch Attack on Object Detectors.....	74
<i>Xin Liu, Huanrui Yang, Ziwei Liu, Linghao Song, Yiran Chen and Hai Li</i>	

Attacks on Machine Learning: Lurking Danger for Accountability	82
<i>Katja Auernhammer, Ramin Tavakoli Kolagari and Markus Zoppelt</i>	
<hr/>	
Short Poster Papers	
<hr/>	
Towards international standards for evaluating machine learning.....	91
<i>Frank Rudzicz, P Alison Paprica and Marta Janczarski</i>	
Counterfactual Explanations of Machine Learning Predictions: Opportunities and Challenges for AI Safety	95
<i>Kacper Sokol and Peter Flach</i>	
Safe Temporal Planning for Urban Driving.....	99
<i>Bence Cserna, William Doyle, Tianyi Gu and Wheeler Ruml</i>	
Linking Artificial Intelligence Principles	103
<i>Yi Zeng, Enmeng Lu and Cunqing Huangfu</i>	
Emergence of Addictive Behaviors in Reinforcement Learning Agents.....	107
<i>Vahid Behzadan, Roman V. Yampolskiy and Arslan Munir</i>	
Temporally Extended Metrics for Markov Decision Processes	111
<i>Philip Amortila, Marc G. Bellemare, Prakash Panangaden and Doina Precup</i>	
AutoMPC: Efficient Multi-Party Computation for Secure and Privacy-Preserving Cooperative Control of Connected Autonomous Vehicles.....	116
<i>Tao Li, Lei Lin and Siyuan Gong</i>	
Security-preserving Support Vector Machine with Fully Homomorphic Encryption	120
<i>Saerom Park, Jaewook Lee, Jung Hee Cheon, Joohee Lee, Jaeyun Kim and Junyoung Byun</i>	
Bamboo: Ball-Shape Data Augmentation Against Adversarial Attacks from All Directions	122
<i>Huanrui Yang, Jingchi Zhang, Hsin-Pai Cheng, Wenhan Wang, Yiran Chen and Hai Li</i>	