

Table of Contents

Invited Talk to the AI Safety Landscape Session

Towards a Framework for Safety Assurance of Autonomous Systems	1
<i>John McDermid, Yan Jia and Ibrahim Habli</i>	

Session 1: Safe Learning

Learning Modular Safe Policies in the Bandit Setting with Application to Adaptive Clinical Trials	8
<i>Hossein Aboutalebi, Doina Precup and Tibor Schuster</i>	
Metric Learning for Value Alignment	15
<i>Andrea Loreggia, Nicholas Mattei, Francesca Rossi and Kristen Brent Venable</i>	

Session 2: Reinforcement Learning Safety

Penalizing side effects using stepwise relative reachability	22
<i>Victoria Krakovna, Laurent Orseau, Miljan Martic and Shane Legg</i>	
Conservative Agency	29
<i>Alexander Turner, Dylan Hadfield-Menell and Prasad Tadepalli</i>	
Detecting Spiky Corruption in Markov Decision Processes	37
<i>Jason Mancuso, Tomasz Kisielewski, David Lindner and Alok Singh</i>	
Modeling AGI Safety Frameworks with Causal Influence Diagrams	44
<i>Tom Everitt, Ramana Kumar, Victoria Krakovna and Shane Legg</i>	

Session 3: Safe Autonomous Vehicles

On the Susceptibility of Deep Neural Networks to Natural Perturbations	51
<i>Mesut Ozdag, Sunny Raj, Steven L. Fernandes, Alvaro Velasquez, Laura Pullum and Sumit Kumar Jha</i>	
Managing Uncertainty of AI-based Perception for Autonomous Systems	57
<i>Maximilian Henne, Adrian Schwaiger and Gereon Weiss</i>	
A Framework for Safety Violation Identification and Assessment in Autonomous Driving .	61
<i>Lukas Heinzmann, Sina Shafaei, Mohd Hafeez Osman, Christoph Segler and Alois Knoll</i>	

Session 4: AI Value Alignment, Ethics and Bias

The Glass Box Approach: Verifying Contextual Adherence to Values	68
<i>Andrea Aler Tubella and Virginia Dignum</i>	
Requisite Variety in Ethical Utility Functions for AI Value Alignment	75
<i>Nadisha-Marie Aliman and Leon Kester</i>	
Slam the Brakes: Perceptions of Moral Decisions in Driving Dilemmas	82
<i>Holly Wilson and Andreas Theodorou</i>	

Understanding Bias in Datasets using Topological Data Analysis 91
Ramya Srinivasan and Ajay Chander

Poster Papers

Computational Strategies for the Trustworthy Pursuit and the Safe Modeling of Probabilistic Maintenance Commitments 98
Qi Zhang, Edmund Durfee and Satinder Singh

Categorizing Wireheading in Partially Embedded Agents 105
Arushi Majha, Sayan Sarkar and Davide Zagami

Adversarial Exploitation of Policy Imitation 112
Vahid Behzadan and William Hsu

The Challenge of Imputation in Explainable Artificial Intelligence Models 119
Muhammad Ahmad, Carly Eckert and Ankur Teredesai

On the importance of system testing for assuring safety of AI systems 123
Franz Wotawa

Towards Empathic Deep Q-Learning 130
Bart Bussmann, Jacqueline Heinerman and Joel Lehman

Watermarking of DRL Policies with Sequential Triggers 137
Vahid Behzadan and William Hsu