Table of Contents

Invited Talk to the AI Safety Landscape Session

Towards a Framework for Safety Assurance of Autonomous Systems ..................... 1
  John McDermid, Yan Jia and Ibrahim Habli

Session 1: Safe Learning

Learning Modular Safe Policies in the Bandit Setting with Application to Adaptive Clinical Trials ................................................................. 8
  Hossein Aboutalebi, Doina Precup and Tibor Schuster

Metric Learning for Value Alignment .............................................................. 15
  Andrea Loreggia, Nicholas Mattei, Francesca Rossi and Kristen Brent Venable

Session 2: Reinforcement Learning Safety

Penalizing side effects using stepwise relative reachability .................................. 22
  Victoria Krakovna, Laurent Orseau, Miljan Martic and Shane Legg

Conservative Agency ...................................................................................... 29
  Alexander Turner, Dylan Hadfield-Menell and Prasad Tadepalli

Detecting Spiky Corruption in Markov Decision Processes .................................. 37
  Jason Mancuso, Tomasz Kisielewski, David Lindner and Alok Singh

Modeling AGI Safety Frameworks with Causal Influence Diagrams ...................... 44
  Tom Everitt, Ramana Kumar, Victoria Krakovna and Shane Legg

Session 3: Safe Autonomous Vehicles

On the Susceptibility of Deep Neural Networks to Natural Perturbations ............... 51
  Mesut Ozdag, Sunny Raj, Steven L. Fernandes, Alvaro Velasquez, Laura Pullum and Sumit Kumar Jha

Managing Uncertainty of AI-based Perception for Autonomous Systems .............. 57
  Maximilian Henne, Adrian Schwager and Gereon Weiss

A Framework for Safety Violation Identification and Assessment in Autonomous Driving . 61
  Lukas Heinzmann, Sina Shafaei, Mohd Hafeez Osman, Christoph Segler and Alois Knoll

Session 4: AI Value Alignment, Ethics and Bias

The Glass Box Approach: Verifying Contextual Adherence to Values ................... 68
  Andrea Aler Tubella and Virginia Dignum

Requisite Variety in Ethical Utility Functions for AI Value Alignment ................... 75
  Nadisha-Marie Aliman and Leon Kester

Slam the Brakes: Perceptions of Moral Decisions in Driving Dilemmas ............ 82
  Holly Wilson and Andreas Theodorou
Understanding Bias in Datasets using Topological Data Analysis

Ramya Srinivasan and Ajay Chander

Poster Papers

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

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

Adversarial Exploitation of Policy Imitation
Vahid Behzadan and William Hsu

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

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

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

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