WAISE 2018

First International Workshop on Artificial Intelligence Safety Engineering

Sept 18th, 2018

Västerås, Sweden

Huascar Espinoza, CEA LIST, France
Orlando Avila-García, Atos, Spain
Rob Alexander, University of York, UK
Andreas Theodorou, University of Bath, UK

SAFECOMP 2018

*"Designing **AI**-based systems for operation in proximity to and/or in collaboration with humans implies that current **safety engineering** and legal mechanisms need to be revisited to ensure that individuals –and their properties– are not harmed and that the desired benefits outweigh the potential unintended consequences."*

Opening Remarks

- As a first version of WAISE, we expect some confusion and clash of perspectives
- Indeed, the main WAISE goal is to bring together those multiple perspectives!
- The AI Safety community must be voraciously interdisciplinary if it is to be useful
- It must look holistically at AI and safety engineering, jointly with the ethical and legal issues, to build trustable intelligent autonomous machines
- Because of this diversity, it's probably possible to harshly criticise any paper here today… Most likely anyone has missed some important issue…
- So, please do be critical, but temper your criticism with constructive discussions!

Programme (Morning)

8:30 – 9:30 > Keynote: Autonomous Vehicle Safety Technical and Social Issues, Prof. Philip Koopman

9:30 – 10:30 > Session 1: **Machine Learning Safety and Reliability** - Chair: Orlando Avila-García

"Boxing Clever": Practical Techniques for Gaining Insights into Training Data and Monitoring Distribution Shift, Rob Ashmore and Matthew Hill

Mitigation of Policy Manipulation Attacks on Deep Q-Networks with Parameter-Space Noise, Vahid Behzadan and Arslan Munir

What is Acceptably Safe for Reinforcement Learning?, John Bragg and Ibrahim Habli

Debate Panel - Paper Discussants: Jin Zhang, Rob Ashmore

10:30 – 11:00 > Coffee Break - Poster Sessions

11:00 – 12:20 > Session 2: **Uncertainty in Automated Driving** - Chair: Timo Latvala

Uncertainty in Machine Learning Applications - A Practice-Driven Classification of Uncertainty, Michael Kläs and Anna Maria Vollmer

Towards a Framework to Manage Perceptual Uncertainty for Safe Automated Driving, Krzysztof Czarnecki and Rick Salay

Design of a Knowledge-Base Strategy for Capability-Aware Treatment of Uncertainties of Automated Driving Systems, Dejiu Chen, Kenneth Östberg, Matthias Becker, Håkan Sivencrona and Fredrik Warg

Uncertainty in Machine Learning: A Safety Perspective on Autonomous Driving, Sina Shafaei, Stefan Kugele, Mohd Hafeez Osman and Alois Knoll

Debate Panel - Paper Discussants: Hari Balaji

12:20 – 13:20 > Lunch

Programme (Afternoon)

13:20 – 13:50 > Invited Talk: Challenges in the Qualification of Safety-Critical Machine Learning-based Components, Prof. François Terrier

13:50 – 14:50 > Session 3: **Challenges in AI Safety** - Chair: Rob Ashmore
Considerations of Artificial Intelligence Safety Engineering for Unmanned Aircraft, Sebastian Schirmer, Christoph Torens, Florian Nikodem and Johann Dauer
Could We Issue Driving Licenses to Autonomous Vehicles?, Jingyue Li, Jin Zhang and Nektaria Kaloudi
Concerns on the differences between AI and system safety mindsets impacting autonomous vehicles safety, Alexandre Moreira Nascimento, Lucio Vismari, Paulo Cugnasca, Joao Camargo Jr., Jorge Almeida Jr., Rafia Inam, Elena Fersman, Alberto Hata and Maria Marquezini
Debate Panel - Paper Discussants: Huascar Espinoza, Alexandre Moreira Nascimento

14:50 – 15:30 > Session 4: **Ethically Aligned Design of Autonomous Systems**- Chair: Rob Alexander
The Moral Responsibility Gap and the Increasing Autonomy of Systems, Zoe Porter, Ibrahim Habli, Helen Monkhouse and John Bragg
Design Requirements for a Moral Machine for Autonomous Weapons, Ilse Verdiesen, Virginia Dignum and Iyad Rahwan
Debate Panel - Paper Discussants: Mauricio Castillo-Effen, Orlando Avila-Garcia

15:30 – 16:00 > Coffee Break - Poster Sessions

16:00 – 17:00 > Session 5: **Human-Inspired Approaches to AI Safety** - Chair: Andreas Theodorou
AI Safety and Reproducibility: Establishing Robust Foundations for the Neuropsychology of Human Values, Gopal Sarma, Nick Hay and Adam Safron
A Psychopathological Approach to Safety Engineering in AI and AGI, Vahid Behzadan, Arslan Munir and Roman Yampolskiy
Why Bad Coffee? Explaining Agent Plans with Valuings, Michael Winikoff, Virginia Dignum and Frank Dignum
Debate Panel - Paper Discussants: Ilse Verdiesen

17:00 – 17:40 > Session 6: **Runtime Risk Assessment in Automated Driving** - Chair: Jérémie Guiochet
Dynamic Risk Assessment for Vehicles of Higher Automation Levels by Deep Learning, Patrik Feth, Mohammed Naveed Akram, René Schuster and Oliver Wasenmüller
Improving Image Classification Robustness using Predictive Data Augmentation, Harisubramanyabalaji Subramani Palanisamy, Shafiq Ur Rehman, Mattias Nyberg and Joakim Gustavsson
Debate Panel - Paper Discussants: Timo Latvala

17:40 – 18:00 > Wrap-up - Best Paper Award

Some Additional Information

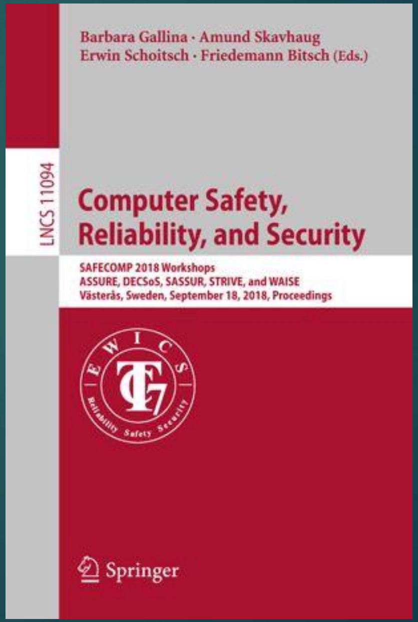
- Voting for WAISE 2018 Best Paper Award:

  www.menti.com – Cod: 88 90 50

- Springer Proceedings are freely available until Oct 16:
  - link at the WAISE website

- Presentations will be available in the website

- We hope you enjoy WAISE 2018!



Barbara Gallina · Amund Skavhaug
Erwin Schoitsch · Friedemann Bitsch (Eds.)

LNCS 11094

Computer Safety, Reliability, and Security

SAFECOMP 2018 Workshops
ASSURE, DECSoS, SASSUR, STRIVE, and WAISE
Västerås, Sweden, September 18, 2018, Proceedings

Springer