www.waise2018.com

WAISE

Sept 18th 2018

First International Workshop on

Artificial Intelligence Safety Engineering

Västerås, Sweden

@SAFECOMP 2018

WHAT is about

WAISE provides a forum for thematic presentations and in-depth discussions on **AI safety** engineering**, ethically aligned design, regulation** and **standards** for **AI-based systems**.

Submit your contribution!

Important DATES [extended]

May 29	Jun 11	Jun 21
Paper Submission	Acceptance Notification	Camera-ready Version

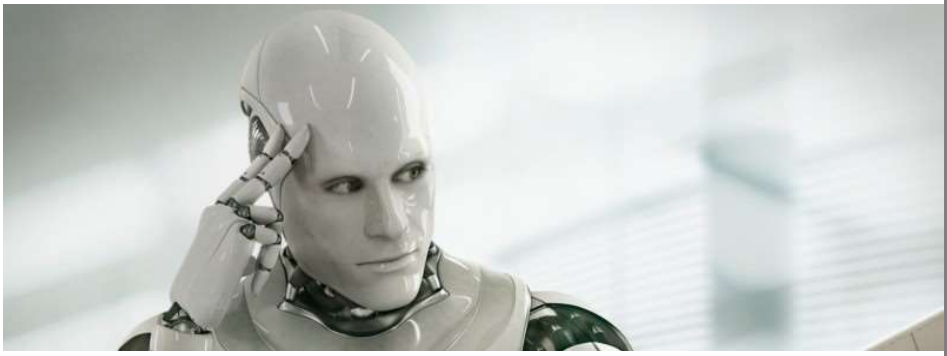
Research, engineering and regulatory frameworks are needed to achieve the full potential of **Artificial Intelligence (AI)** as they will guarantee a standard level of safety and settle issues such as compliance with ethical standards and liability for accidents involving e.g., autonomous cars. Designing 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.

The different approaches taken to **AI safety** go from pure theoretical (moral philosophy or ethics) to pure practical (engineering) planes. It appears as essential to combine philosophy and theoretical science with applied science and engineering in order to create safe machines. This should become an interdisciplinary approach covering technical (engineering) aspects of how to actually create, test, deploy, operate and evolve safe **AI-based systems**, as well as broader strategic, ethical and policy issues.

Topics

Contributions are sought in (but are not limited to) the following topics:

- Avoiding negative side effects
- Safety in AI-based system architectures: safety by design
- Runtime monitoring and (self-)adaptation of AI safety
- Safe machine learning and meta-learning
- Safety constraints and rules in decision making systems
- Continuous Verification and Validation (V&V) of safety properties
- AI-based system predictability
- Model-based engineering approaches to AI safety
- Ethically aligned design of AI-based systems
- Machine-readable representations of ethical principles and rules
- The values alignment problem
- The goals alignment problem
- Accountability, responsibility and liability of AI-based systems
- Uncertainty in AI
- AI safety risk assessment and reduction
- Loss of values and the catastrophic forgetting problem
- Confidence, self-esteem and the distributional shift problem
- Reward hacking and training corruption
- Weaponization of AI-based systems
- Self-explanation, self-criticism and the transparency problem
- Simulation for safe exploration and training
- Human-machine interaction safety
- AI applied to safety engineering
- Zero-sum and the trolley problem
- Regulating AI-based systems: safety standards and certification
- Human-in-the-loop and the scalable oversight problem
- Algorithmic bias and AI discrimination
- AI safety education and awareness
- Experiences in AI-based safety-critical systems, including manufacturing, health, transport, robotics, critical infrastructures, among others



Submissions

Format:

- ❑ Scientific Paper: 12 pages (PDF, Springer LNCS)
- ❑ Position Paper: 6 pages (PDF, Springer LNCS)
- ❑ Talk/Session proposal: abstract (PDF)

Further Information

Website: http://www.waise2018.com
Contact: waise2018@easychair.org
Submission link:
https://easychair.org/conferences/?conf=waise2018

Organization Committee

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

Steering Committee

Stuart Russell, UC Berkeley, USA
Raja Chatila, ISIR - Sorbonne University, France
Roman V. Yampolskiy, University of Louisville, USA
Nozha Boujemaa, DATAIA Institute & INRIA, France
Mark Nitzberg, Center for Human-Compatible AI, USA
Philip Koopman, Carnegie Mellon University, USA

Programme Committee

Roman V. Yampolskiy, University of Louisville, USA
Stuart Russell, UC Berkeley, USA
Raja Chatila, ISIR - Sorbonne University, France
Nozha Boujemaa, DATAIA Institute & INRIA, France
Mark Nitzberg, Center for Human-Compatible AI, USA
Victoria Krakovna, Google DeepMind, UK
Chokri Mraidha, CEA LIST, France
Heather Roff, Leverhulme Centre for the Future of Intelligence, UK
Bernhard Kaiser, ANSYS, Germany
John Favaro, INTECS, Italy
Jonas Nilsson, Zenuity, Sweden
Philippa Ryan, Adelard, UK
José Hernández-Orallo, Universitat Politècnica de València, Spain
Andrew Banks, LDRA, UK
Carlos Hernández, TU Delft, Netherlands
José M. Faria, Safe Perspective Ltd., UK
Philip Koopman, Carnegie Mellon University, USA
Florent Kirchner, CEA LIST, France
Joanna Bryson, University of Bath, UK
Stefan Kugele, Technical University of Munich, Germany
Virginia Dignum, TU Delft, Netherlands
Timo Latvala, Space Systems Finland, Finland
Mehrdad Saadatmand, RISE SICS, Sweden
Rick Salay, University of Waterloo, Canada
Lavinia Burski, AECOM, UK
Jérémie Guiochet, LAAS-CNRS, France
Mario Gleirscher, University of York, UK
François Terrier, CEA LIST, France
Rob Ashmore, Defence Science & Technology Laboratory, UK
Erwin Schoitsch, Austrian Institute of Technology, Austria
Chris Allsopp, Frazer-Nash Consultancy, UK
Mauricio Castillo-Effen, Lockheed Martin, USA