Call for Contributions

First International Workshop on Artificial Intelligence Safety Engineering (WAISE 2018)

In conjunction with SAFECOMP 2018

Västerås, Sweden, Sept. 18, 2018

http://www.waise2018.com

SCOPE

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

Increasing levels of AI in "smart" sensory-motor loops allow intelligent systems to perform in increasingly dynamic uncertain complex environments with increasing degrees of **autonomy**, with human being progressively ruled out from the control loop. Adaptation to the environment is being achieved by **Machine Learning (ML)** methods rather than more traditional engineering approaches, such as system modelling and programming. Recently, certain ML methods are proving themselves specially promising, such as deep learning, reinforcement learning and their combination. However, the **inscrutability** or opaqueness of the statistical models for perception and decision-making we build through them pose yet another challenge. Moreover, the combination of autonomy and inscrutability in these AI-based systems is particularly challenging in safety-critical applications, such as autonomous vehicles, personal care or assistive robots and collaborative industrial robots.

The **WAISE workshop** is intended to explore new ideas on safety engineering for AI-based systems, ethically aligned design, regulation and standards for AI-based systems. In particular, WAISE will provide a forum for thematic presentations and in-depth discussions about safe AI architectures, bounded morality, ML safety, safe human-machine interaction and safety considerations in automated decision making systems, in a way that makes AI-based systems more trustworthy, accountable and ethically aligned.

WAISE aims at bringing together experts, researchers, and practitioners, from diverse communities, such as AI, safety engineering, ethics, standardization and certification, robotics, cyber-physical systems, safety-critical systems, and application domain communities such as **automotive, healthcare, manufacturing, agriculture, aerospace, critical infrastructures, and retail**.

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 industrial processes, health, automotive systems, robotics, critical infrastructures, among others

IMPORTANT DATES [EXTENDED]

- **Full paper submission**: May 29, 2018
- **Notification of acceptance**: June 11, 2018
- **Camera-ready submission**: June 21, 2018
- **Workshop**: Sept 18, 2018

SUBMISSION AND SELECTION

You are invited to submit **short position papers** (max. 6 pages), full **scientific contributions** (max. 12 pages) or proposals of technical **talk/sessions** (short abstracts). Manuscripts must be submitted as PDF files via **EasyChair** online submission system:

https://easychair.org/conferences/?conf=waise2018

Workshop proceedings will be provided as complementary book to the SAFECOMP Proceedings in Springer LNCS. Please keep your paper format according to SPRINGER LNCS style guidelines:

http://www.springer.com/computer/lncs?SGWID=0-164-6-793341-0

Papers will be peer-reviewed by the Program Committee (minimum 3 reviewers per paper).

For any question, please send an email to: waise2018@easychair.org

COMMITTEES

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 (look at the website: http://www.waise2018.com)