==========

Call for Contributions
The AAAI-19 Workshop on Artificial Intelligence Safety (SafeAI 2019)
Honolulu, Hawaii, Jan 27, 2019
Submission Deadline: Nov 5, 2018
http://www.safeai2019.org

==========

Scope

Safety in Artificial Intelligence (AI) should not be an option, but a design principle. However, there are different levels of safety, different ethical standards and values, and different degrees of liability, for which we face trade-offs or alternative solutions. These choices can only be analyzed holistically if we integrate the technological and the ethical perspectives into the engineering problem, and consider both the theoretical and practical challenges for AI safety. This view must cover a wide range of AI paradigms, including systems that are specific for a particular application, and also those that are more general, and can lead to unanticipated potential risks. We must also bridge short-term with long-term issues, idealistic with pragmatic solutions, operational with policy issues, and industry with academia, to really build, evaluate, deploy, operate and maintain AI-based systems that are truly safe.

This workshop seeks to explore new ideas on AI safety with particular focus on addressing the following questions:

- What is the status of existing approaches in ensuring AI and Machine Learning (ML) safety and what are the gaps?
- How can we engineer trustable AI software architectures?
- How can we make AI-based systems more ethically aligned?
- What safety engineering considerations are required to develop safe human-machine interaction?
- What AI safety considerations and experiences are relevant from industry?
- How can we characterize or evaluate AI systems according to their potential risks and vulnerabilities?
- How can we develop solid technical visions and new paradigms about AI Safety?
- How do metrics of capability and generality, and the trade-offs with performance affect safety?

The main interest of the proposed workshop is to look holistically at AI and safety engineering, jointly with the ethical and legal issues, to build trustable intelligent autonomous machines.

Topics

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

- Safety in AI-based system architectures
- Continuous V&V and predictability of AI safety properties
- Runtime monitoring and (self-)adaptation of AI safety
- Accountability, responsibility and liability of AI-based systems
- Effect of uncertainty in AI safety
- Avoiding negative side effects in AI-based systems
- Role and effectiveness of oversight: corrigibility and interruptibility
- Loss of values and the catastrophic forgetting problem
- Confidence, self-esteem and the distributional shift problem
- Safety of Artificial General Intelligence (AGI) systems and the role of generality
- Reward hacking and training corruption
- Self-explanation, self-criticism and the transparency problem

- Human-machine interaction safety
- Regulating AI-based systems: safety standards and certification
- Human-in-the-loop and the scalable oversight problem
- Evaluation platforms for AI safety
- AI safety education and awareness
- Experiences in AI-based safety-critical systems, including industrial processes, health, automotive systems, robotics, critical infrastructures, among others

Format

To deliver a truly memorable event, we will follow a highly interactive format that will include invited talks and thematic sessions. The thematic sessions will be structured into short pitches and a common panel slot to discuss both individual paper contributions and shared topic issues. Three specific roles are part of this format: session chairs, presenters and paper discussants. The workshop will be organized as a full day meeting.

Attendance is open to all. At least one author of each accepted submission must be present at the workshop.

Submissions

You are invited to submit short position papers (2-4 pages), full technical papers (6-8 pages) or proposals of technical talk (up one-page abstract). Manuscripts must be submitted as PDF files via EasyChair online submission system: https://easychair.org/conferences/?conf=SafeAI2019

Please keep your paper format according to AAAI Formatting Instructions (two-column format). The AAAI author kit can be downloaded from: https://www.aaai.org/Publications/Templates/AuthorKit18.zip

Papers will be peer-reviewed by the Program Committee (2-3 reviewers per paper). The workshop follows a single-blind reviewing process. However, we will also accept anonymized submissions.

For any question, please send an email to: safeai2019@easychair.org

Organization Committee

Huáscar Espinoza, CEA LIST, France
Seán Ó hÉigeartaigh, University of Cambridge, UK
Xiaowei Huang, University of Liverpool, UK
José Hernández-Orallo, Universitat Politècnica de València, Spain

Program Committee

Look at the website: http://www.safeai2019.org