The AAAI-19 Workshop on Artificial Intelligence Safety (SafeAI 2019)

Huáscar Espinoza[1], Seán Ó hÉigeartaigh[2], Xiaowei Huang[3],
José Hernández-Orallo[4] and Mauricio Castillo-Effen[5]

[1] CEA LIST, Gif-sur-Yvette, France
huascar.espinoza@cea.fr

[2] University of Cambridge, Cambridge, United Kingdom
so348@cam.ac.uk

[3] University of Liverpool, Liverpool, United Kingdom
xiaowei.huang@liverpool.ac.uk

[4] Universitat Politècnica de València, Valencia, Spain
jorallo@dsic.upv.es

[5] Lockheed Martin, Advanced Technology Laboratories, Arlington, VA, USA
mauricio.castillo-effen@lmco.com

Abstract

This preface introduces the AAAI-19 Workshop on Artificial Intelligence Safety (SafeAI 2019), held at the Thirty-Third AAAI Conference on Artificial Intelligence on January 27, 2019 in Honolulu, Hawaii, USA.

Introduction

Safety in Artificial Intelligence (AI) should not be an option, but a design principle. However, there are varying levels of safety, diverse sets of ethical standards and values, and varying degrees of liability, for which we need to deal with trade-offs or alternative solutions. These choices can only be analyzed holistically if we integrate the technological and 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, considering systems that are specific for a particular application, and also those that are more general, which may lead to unanticipated risks. We must bridge the short-term with the long-term perspectives, idealistic with pragmatic solutions, operational with policy issues, and industry with academia, to build, evaluate, deploy, operate and maintain AI-based systems that are truly safe.

The AAAI-19 Workshop on Artificial Intelligence Safety (SafeAI 2019) 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 trade-offs with performance affect safety?

The main interest of SafeAI 2019 is to look holistically at AI and safety engineering, jointly with the ethical and legal issues, to build trustable intelligent autonomous machines. The first edition of SafeAI was held in January 27, 2019, in Honolulu, Hawaii (USA) as part of the Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19).

Program

The Program Committee (PC) received 33 submissions, in the following categories:

- Short position papers – 13 submissions.
- Full scientific contributions – 18 submissions.
- Proposals of technical talks – 2 submissions.

Each of the papers was peer-reviewed by at least two PC members, by following a single-blind reviewing process. The committee decided to accept 12 papers (4 position papers and 8 scientific papers) and 1 talk, resulting in an overall acceptance rate of 39%. We additionally invited two talks, which were not submitted to the call, and accepted 10 submissions as short papers for poster presentation.

The SafeAI 2019 program was organized in five thematic sessions, two keynote and three (invited) talks.

The thematic sessions followed a highly interactive format. They were structured into short pitches and a common panel slot to discuss both individual paper contributions and shared topic issues. Three specific roles were part of this format: session chairs, presenters and session discussants.

- *Session Chairs* introduced sessions and participants. The Chair moderated session and plenary discussions, took care of the time, and gave the word to speakers in the audience during discussions.
- *Presenters* gave a paper pitch in 10 minutes and then participated in the debate slot.
- *Session Discussants* prepared the discussion of individual papers and the plenary debate. The discussant gave a critical review of the session papers.

The mixture of topics has been carefully balanced, as follows:

Session 1: Safe Planning and Operation of Autonomous Systems

- Minimizing the Negative Side Effects of Planning with Reduced Models, Sandhya Saisubramanian and Shlomo Zilberstein.
- Robust Motion Planning and Safety Benchmarking in Human Workspaces, Shih-Yun Lo, Shani Alkoby and Peter Stone.
- Enter the Matrix: Safely Interruptible Autonomous Systems via Virtualization, Mark Riedl and Brent Harrison.

Session 2: New Paradigms in AI and AGI Safety

- Towards Robust End-to-End Alignment, Lê Nguyên Hoang.
- Integrative Biological Simulation, Neuropsychology, and AI Safety, Gopal Sarma, Adam Safron and Nick Hay.

Session 3: Safety in Automated Driving

- How Many Operational Design Domains, Objects, and Events?, Philip Koopman and Frank Fratrik.
- Monitoring Safety of Autonomous Vehicles with Crash Prediction Networks, Saasha Nair, Sina Shafaei, Stefan Kugele, Mohd Hafeez Osman and Alois Knoll.

Session 4: Safety-Related AI Requirements and Characteristics

- Requirements Assurance in Machine Learning, Alec Banks and Rob Ashmore.
- Surveying Safety-relevant AI Characteristics, Jose Hernandez-Orallo, Fernando Martínez-Plumed, Shahar Avin and Seán Ó hÉigeartaigh.

Session 5: Adversarial Machine Learning

- Detecting Backdoor Attacks on Deep Neural Networks by Activation Clustering, Bryant Chen, Wilka Carvalho, Nathalie Baracaldo, Heiko Ludwig, Benjamin Edwards, Taesung Lee, Ian Molloy and Biplav Srivastava.
- DPATCH: An Adversarial Patch Attack on Object Detectors, Xin Liu, Huanrui Yang, Ziwei Liu, Linghao Song, Yiran Chen and Hai Li.
- Attacks on Machine Learning: Lurking Danger for Accountability, Katja Auernhammer, Ramin Tavakoli Kolagari and Markus Zoppelt

Additionally, SafeAI was proud to bring great inspirational speakers:

Keynotes

- Dr. Sandeep Neema (DARPA), Assured Autonomy.
- Prof. Francesca Rossi (IBM and University of Padova), Ethically Bounded AI.

Invited Talks

- Dr. Peter Eckersley (Partnership on AI), Impossibility and Uncertainty Theorems in AI Value Alignment (or why your AGI should not have a utility function).
- Dr. Ian Goodfellow (Google Brain), Adversarial Robustness for AI Safety.
- Prof. Alessio R. Lomuscio (Imperial College London), Reachability Analysis for Neural Agent-Environment Systems.

Posters were presented in 2-minute pitches and most of them are also part of this volume as short papers.

Posters

- Towards international standards for evaluating machine learning, Frank Rudzicz, P Alison Paprica and Marta Janczarski.
- Counterfactual Explanations of Machine Learning Predictions: Opportunities and Challenges for AI Safety, Kacper Sokol and Peter Flach.
- Safe Temporal Planning for Urban Driving, Bence Cserna, William Doyle, Tianyi Gu and Wheeler Ruml.
- Linking Artificial Intelligence Principles, Yi Zeng, Enmeng Lu and Cunqing Huangfu.
- Emergence of Addictive Behaviors in Reinforcement Learning Agents, Vahid Behzadan, Roman V. Yampolskiy and Arslan Munir.
- Temporally Extended Metrics for Markov Decision Processes, Philip Amortila, Marc G. Bellemare, Prakash Panangaden and Doina Precup.
- Exploring Interfaces to Democratize AI Constraint Generation, Travis Mandel, Jahnu Best, Randall Tanaka, Hiram Temple, Chansen Haili and Roy Szeto [*no paper*].
- AutoMPC: Efficient Multi-Party Computation for Secure and Privacy-Preserving Cooperative Control of Connected Autonomous Vehicles, Tao Li, Lei Lin and Siyuan Gong.
- Security-preserving Support Vector Machine with Fully Homomorphic Encryption, Saerom Park, Jaewook Lee, Jung Hee Cheon, Juhee Lee, Jaeyun Kim and Junyoung Byun.
- Bamboo: Ball-Shape Data Augmentation Against Adversarial Attacks from All Directions, Huanrui Yang, Jingchi Zhang, Hsin-Pai Cheng, Wenhan Wang, Yiran Chen and Hai Li.

Acknowledgements

We thank all those who submitted papers to SafeAI 2019 and congratulate the authors whose papers and posters were selected for inclusion into the workshop program and proceedings.

We specially thank our distinguished PC members, for reviewing the submissions and providing useful feedback to the authors:

- Stuart Russell, UC Berkeley, USA
- Francesca Rossi, IBM and University of Padova, Italy
- 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
- François Terrier, CEA LIST, France
- Orlando Avila-García, Atos, Spain
- Kobi Leins, Australian National University, Australia
- Rob Alexander, University of York, UK
- Peter Eckersley, Electronic Frontier Foundation, USA
- Andreas Theodorou, University of Bath, UK
- Philip Koopman, Carnegie Mellon University, USA
- Chokri Mraidha, CEA LIST, France
- Heather Roff, Google DeepMind, UK
- Bernhard Kaiser, ANSYS, Germany
- Brent Harrison, University of Kentucky, USA
- John Favaro, INTECS, Italy
- Rob Ashmore, Defence Science and Technology Laboratory, UK
- Martin Vechev, ETH Zurich, Switzerland
- Jonas Nilsson, Zenuity, Sweden
- Philippa Ryan, Adelard, UK
- Florent Kirchner, CEA LIST, France
- Andrew Banks, LDRA, UK
- Stefan Kugele, Technical University of Munich, Germany
- Marta Kwiatkowska, University of Oxford, UK
- Virginia Dignum, TU Delft, Netherlands
- Yi Zeng, Institute of Automation, Chinese Academy of Sciences, China
- Timo Latvala, Space Systems Finland, Finland
- Mehrdad Saadatmand, RISE SICS, Sweden
- Rick Salay, University of Waterloo, Canada
- Jérémie Guiochet, LAAS-CNRS, France
- Mario Gleirscher, University of York, UK
- Chris Allsopp, Frazer-Nash Consultancy, UK
- Daniela Cancila, CEA LIST, France
- 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
- Mauricio Castillo-Effen, Lockheed Martin, USA

As well as the additional reviewers:

- Zac Hatfield-Dodds, Australian National University, Australia
- Patrik Hoyer, Space Systems Finland, Finland
- Paolo Panaroni, Intecs, Italy
- Renaud Sirdey, CEA LIST, France

We thank Dr. Sandeep Neema, Prof. Francesca Rossi, Dr. Peter Eckersley, Dr. Ian Goodfellow, and Prof. Alessio R. Lomuscio for their interesting talks on the current challenges of safety and ethics for AI and autonomy.

Finally, yet importantly, we thank the AAAI-19 organization for providing an excellent framework for SafeAI 2019.