The IJCAI-19 Workshop on Artificial Intelligence Safety (AISafety 2019)

Huáscar Espinoza[1], Han Yu[2], Xiaowei Huang[3], Freddy Lecue[4], Cynthia Chen[5], José Hernández-Orallo[6], Seán Ó hÉigeartaigh[7], and Richard Mallah[8]
[1]Commissariat à l´Energie Atomique, France
[2]Nanyang Technological University, Singapore
[3]University of Liverpool, UK
[4]Thales, Canada
[5]University of Hong Kong, China
[6]Universitat Politècnica de València, Spain
[7]University of Cambridge, UK
[8]Future of Life Institute, USA

Abstract

This preface introduces the IJCAI-19 Workshop on Artificial Intelligence Safety (AISafety 2019), held at the 28th International Joint Conference on Artificial Intelligence (IJCAI) on August 11-12, 2019 in Macao, China.

1 Introduction

In the last decade, there has been a growing concern on risks of Artificial Intelligence (AI). Safety is becoming increasingly relevant as humans are progressively side-lined from the decision/control loop of intelligent and learning-enabled machines. In particular, the technical foundations and assumptions on which traditional safety engineering principles are based, are inadequate for systems in which AI algorithms, and in particular Machine Learning (ML) algorithms, are interacting with people and/or the environment at increasingly higher levels of autonomy. We must also consider the connection between the safety challenges posed by present-day AI systems, and more forward-looking research focused on more capable future AI systems, up to and including Artificial General Intelligence (AGI).

The IJCAI-19 Workshop on Artificial Intelligence Safety (AISafety 2019) seeks to explore new ideas on AI safety with particular focus on addressing the following questions:

- How can we engineer trustable AI software architectures?
- Do we need to specify and use bounded morality in system engineering to make AI-based systems more ethically aligned?
- What is the status of existing approaches in ensuring AI and ML safety and what are the gaps?
- What safety engineering considerations are required to develop safe human-machine interaction in automated decision-making systems?
- What AI safety considerations and experiences are relevant from industry?
- How can we characterise or evaluate AI systems according to their potential risks and vulnerabilities?
- How can we develop solid technical visions and paradigm shift articles about AI Safety?
- How do metrics of capability and generality affect the level of risk of a system and how trade-offs can be found with performance?
- How do AI system feature for example ethics, explainability, transparency, and accountability relate to, or contribute to, its safety?
- How to evaluate AI safety?

The main interest of AISafety 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 AISafety was held in August 11-12, 2019, in Macao (China), as part of the 28th International Joint Conference on Artificial Intelligence (IJCAI-19). The AISafety workshop is organized as a "sister workshop" to other two workshops: WAISE (https://www.waise.org/) and to SafeAI (http://www.safeai2019.org).

As part of this IJCAI workshop, we also started the ***AI Safety Landscape*** initiative. This initiative aims at defining an AI safety landscape providing a "view" of the current needs, challenges and state of the art and the practice of this field. Further information about this initiative can be found at: https://www.ai-safety.org/ai-safety-landscape.

2 Programme

The Programme Committee (PC) received 36 submissions, in the following categories:

- Short position papers – 9 submissions.
- Full scientific contributions – 23 submissions.
- Proposals of technical talks – 4 submissions.

Each of the papers was peer-reviewed by at least three PC members, by following a single-blind reviewing process. The committee decided to accept 13 papers (2 position papers and 11 scientific papers) and 2 talks, resulting in an overall acceptance rate of 42%. We additionally invited 1 talk, which was not submitted to the call, and accepted 7 submissions as short papers for poster presentation.

AISafety 2019 has been planned as a two-day workshop with general AI Safety topics in the first day and AI Safety Landscape talks and discussions during the second day.

2.1. First Workshop Day (Aug 11)

The AISafety 2019 programme on Aug 11 was organized in four thematic sessions, one keynote and two invited talks.

The thematic sessions followed a highly interactive format. They were structured into short talks 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 talk 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 Learning

- Learning Modular Safe Policies in the Bandit Setting with Application to Adaptive Clinical Trials. Hossein Aboutalebi, Doina Precup and Tibor Schuster.
- Metric Learning for Value Alignment. Andrea Loreggia, Nicholas Mattei, Francesca Rossi and Kristen Brent Venable.

Session 2: Reinforcement Learning Safety

- Penalizing side effects using stepwise relative reachability. Victoria Krakovna, Laurent Orseau, Miljan Martic and Shane Legg.
- Conservative Agency. Alexander Turner, Dylan Hadfield-Menell and Prasad Tadepalli.
- Detecting Spiky Corruption in Markov Decision Processes. Alok Singh, Jason Mancuso, David Lindner and Tomasz Kisielewski. Detecting Spiky Corruption in Markov Decision Processes.
- Modeling AGI Safety Frameworks with Causal Influence Diagrams. Tom Everitt, Ramana Kumar, Victoria Krakovna and Shane Legg.

Session 3: Safe Autonomous Vehicles

- On the Susceptibility of Deep Neural Networks to Natural Perturbations. Mesut Ozdag, Sunny Raj, Steven L. Fernandes, Alvaro Velasquez, Laura Pullum and Sumit Kumar Jha.
- Managing Uncertainty of AI-based Perception for Autonomous Systems. Maximilian Henne, Adrian Schwaiger and Gereon Weiss.
- A Framework for Safety Violation Identification and Assessment in Autonomous Driving. Lukas Heinzmann, Sina Shafaei, Mohd Hafeez Osman, Christoph Segler and Alois Knoll.

Session 4: AI Value Alignment, Ethics and Bias

- The Glass Box Approach: Verifying Contextual Adherence to Values. Andrea Aler Tubella and Virginia Dignum.
- Requisite Variety in Ethical Utility Functions for AI Value Alignment. Nadisha-Marie Aliman and Leon Kester
- Slam the Brakes: Perceptions of Moral Decisions in Driving Dilemmas. Holly Wilson, Andreas Theodorou and Joanna Bryson.
- Understanding Bias in Datasets using Topological Data Analysis. Ramya Srinivasan and Ajay Chander.

Additionally, AISafety was proud to bring great inspirational speakers:

Keynote

- Joel Lehman (Uber AI Labs, USA), AI Safety for Evolutionary Computation, Evolutionary Computation for AI Safety.

Invited Talks

- Shlomo Zilberstein (University of Massachusetts Amherst, USA), AI Safety Based on Competency Models.
- Yang Liu (WeBank, China), User Privacy, Data Confidentiality and AI Safety in Collaborative Learning.

Posters were presented in 2-minutes pitches and are also part of this volume as poster papers.

Posters

- Computational Strategies for the Trustworthy Pursuit and the Safe Modeling of Probabilistic Maintenance Commitments. Qi Zhang, Edmund Durfee and Satinder Singh
- Categorizing Wireheading in Partially Embedded Agents. Arushi Majha, Sayan Sarkar and Davide Zagami
- Adversarial Exploitation of Policy Imitation. Vahid Behzadan and William Hsu.
- The Challenge of Imputation in Explainable Artificial Intelligence Models. Muhammad Ahmad, Carly Eckert and Ankur Teredesai
- On the importance of system testing for assuring safety of AI systems. Franz Wotawa
- Towards Empathic Deep Q-Learning. Bart Bussmann, Jacqueline Heinerman and Joel Lehman
- Watermarking of DRL Policies with Sequential Triggers. Vahid Behzadan and William Hsu.

2.2. Second Workshop Day (Aug 12): Landscape

The second-day workshop (AI Safety Landscape) sessions on August 12 were organized into by-invitation talks and panels with structured discussions. The by-invitation talks focused on diverse topics contributing to understand the AI Safety Landscape, in terms of their scientific and technical challenges, industrial and academic opportunities, as well as gaps and pitfalls.

AISafety 2019 was proud to bring great industry, academic and research leaders as invited speakers:

Invited Talks

- Richard Mallah (Future of Life Institute, USA): Creating a Deep Model of AI Safety Research.
- John McDermid (University of York, UK): Towards a Framework for Safety Assurance of Autonomous Systems [*published as part of this Proceedings volume*].
- Gopal Sarma (Broad Institute of MIT and Harvard, USA): AI Safety and The Life Sciences.
- Xiaowei Huang (University of Liverpool, UK): Formal Methods in Certifying Learning-Enabled Systems.
- Virginia Dignum (University of Umeå, Sweden): AI Safety for Humans.
- Raja Chatila (Sorbonne University, France): Towards Trustworthy Autonomous and Intelligent Systems.
- Jeff Cao (Tencent Research Institute, China): AI Principles and Ethics by Design.
- Victoria Krakovna (Google DeepMind, UK): Specification, Robustness and Assurance Problems in AI Safety.

One important ambition of this initiative is to align and synchronize the proposed activities and outcomes with other related initiatives. This AI Safety Landscape work will follow up with future meetings and workshops.

3 Acknowledgements

We thank all those who submitted papers to AISafety 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
- Victoria Krakovna, Google DeepMind, UK
- Peter Eckersley, Partnership on AI, USA
- Riccardo Mariani, Intel, Italy
- Brent Harrison, University of Kentucky, USA
- Siddartha Khastgir, University of Warwick, UK
- Emmanuel Arbaretier, Apsys-Airbus, France
- Martin Vechev, ETH Zurich, Switzerland
- Sandhya Saisubramanian, University of Massachusetts Amherst, USA
- Alessio R. Lomuscio, Imperial College London, UK
- Mauricio Castillo-Effen, Lockheed Martin, USA
- Yi Zeng, Chinese Academy of Sciences, China
- Brian Tse, Affiliate at University of Oxford, China
- Sandeep Neema, DARPA, USA
- Michael Paulitsch, Intel, Germany
- Elizabeth Bondi, University of Southern California, USA
- Hélène Waeselynck, CNRS LAAS, France
- Rob Alexander, University of York, UK
- Vahid Behzadan, Kansas State University, USA
- Simon Fürst, BMW, Germany
- Chokri Mraidha, CEA LIST, France
- Fuxin Li, Oregon State University, USA
- Francesca Rossi, IBM and University of Padova, Italy
- Ian Goodfellow, Google Brain, USA
- Yang Liu, Webank, China
- Ramana Kumar, Google DeepMind, UK
- Javier Ibañez-Guzman, Renault, France
- Dragos Margineantu, Boeing, USA
- Joanna Bryson, University of Bath, UK
- Heather Roff, Johns Hopkins University, USA
- Raja Chatila, Sorbonne University, France
- Hang Su, Tsinghua University, China
- François Terrier, CEA LIST, France
- Guy Katz, Hebrew University of Jerusalem, Israel
- Alec Banks, Defence Science and Technology Laboratory, UK
- Gopal Sarma, Emory University, USA
- Lê Nguyên Hoang, EPFL, Switzerland
- Roman Nagy, BMW, Germany
- Nathalie Baracaldo, IBM Research, USA
- Toshihiro Nakae, DENSO Corporation, Japan
- Peter Flach, University of Bristol, UK
- Richard Cheng, California Institute of Technology, USA

- José M. Faria, Safe Perspective, UK
- Ramya Ramakrishnan, Massachusetts Institute of Technology, USA
- Gereon Weiss, Fraunhofer ESK, Germany
- Huáscar Espinoza, Commissariat à l´Energie Atomique, France
- Han Yu, Nanyang Technological University, Singapore
- Xiaowei Huang, University of Liverpool, UK
- Freddy Lecue, Thales, Canada
- Cynthia Chen, University of Hong Kong, China
- José Hernández-Orallo, Universitat Politècnica de València, Spain
- Seán Ó hÉigeartaigh, University of Cambridge, UK
- Richard Mallah, Future of Life Institute, USA

As well as the additional reviewers:

- George Amariucai, Kansas State University, USA
- Neale Ratzlaff, Oregon State University, USA

We thank Joel Lehman, Shlomo Zilberstein, Yang Liu, Richard Mallah, John McDermid, Gopal Sarma, Xiaowei Huang, Virginia Dignum, Raja Chatila, Jeff Cao and Victoria Krakovna for their interesting talks on the current challenges of AI safety.

We would like to specially thank our sponsors, which funded the Best Paper Award and the video-recording of the AI Safety Landscape sessions:

- Assuring Autonomy International Programme (AAIP).
- Partnership on AI.
- The Centre for the Study of Existential Risk (CSER).

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