The IJCAI-PRICAI-20 Workshop on Artificial Intelligence Safety (AISafety 2020)

Huáscar Espinoza[1], John McDermid[2], Xiaowei Huang[3], Mauricio Castillo-Effen[4], Xin Cynthia Chen[5], José Hernández-Orallo[6], Seán Ó hÉigeartaigh[7], and Richard Mallah[8]

[1]Commissariat à l´Energie Atomique, France
[2]Univeristy of York, UK
[3]University of Liverpool, UK
[4]Loockheed Martin, USA
[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 Second Workshop on Artificial Intelligence Safety (AISafety 2020), held at the 29th International Joint Conference on Artificial Intelligence and the 17th Pacific Rim International Conference on Artificial Intelligence (IJCAI-PRICAI) in January 2021, Japan.

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-PRICAI-20 Workshop on Artificial Intelligence Safety (AISafety 2020) seeks to explore new ideas on AI safety with particular focus on addressing the following questions:

- How can we engineer trustworthy AI system architectures?
- Do we need to specify and use bounded morality in engineering more ethically-aligned AI-based systems?
- What is the status of existing approaches in ensuring AI and ML safety and what are the gaps?
- How to evaluate AI safety?
- What AI safety considerations and experiences are relevant from industry?
- What safety engineering considerations are required to develop safe human-machine interaction in automated decision-making systems?
- 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?

The main interest of AISafety 2020 is to look holistically at AI and safety engineering, jointly with the ethical and legal issues, to build trustable intelligent autonomous machines. The second edition of AISafety will be held in January 2021, in Japan, as part of the 29th International Joint Conference on Artificial Intelligence and the 17th Pacific Rim International Conference on Artificial Intelligence (IJCAI-PRICAI). The AISafety workshop is organized as a "sister workshop" to two other workshops: WAISE[1] and SafeAI[2].

[1] https://www.waise.org
[2] http://www.safeaiw.org

As part of this workshop, we also host discussions related to the ***AI Safety Landscape*** initiative[3]. 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.

2 Programme

The Programme Committee (PC) received 25 submissions, in the following categories:

- Short position papers – 4 submission.
- Full scientific contributions – 20 submissions.
- Proposals of technical talks – 1 submission.

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 11 papers and 1 talk, resulting in an overall acceptance rate of 48%. We additionally accepted 6 submissions as poster presentations, (5 of which are included in this proceedings, as poster papers.

AISafety 2020 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. Since the workshop has been delayed, together with IJCAI-PRICAI-20, from July 2020 to January 2021, due to the COVID-19 pandemic, we do not have yet the full list of invited talks for the first day and no specific talk allocated to the second day. The exact date and format (in-person or virtual conference) is still under discussion by IJCAI-PRICAI organizers at the date of publication of AISafety-20 Proceedings.

2.1. First Workshop Day

The AISafety 2020 programme will be organized in four thematic sessions, one keynote and at least three invited talks.

The thematic sessions will be structured into short talks 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 session discussants.

- *Session Chairs* introduce sessions and participants. The Chair moderates session and plenary discussions, takes care of the time, and gives the word to speakers in the audience during discussions.
- *Presenters* give a paper talk in 10 minutes and then participate in the debate slot.
- *Session Discussants* prepare the discussion of individual papers and the plenary debate. The discussant gives a critical review of the session papers.

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

Session 1: Adversarial Machine Learning

- Understanding the One Pixel Attack: Propagation Maps and Locality Analysis. Danilo Vasconcellos Vargas and Jiawei Su.
- Error-Silenced Quantization: Bridging Robustness and Compactness. Zhicong Tang, Yinpeng Dong and Hang Su.
- Evolving Robust Neural Architectures to Defend from Adversarial Attacks. Shashank Kotyan and Danilo Vasconcellos Vargas.

Session 2: AI Safety Landscape

- Update Report: AI Safety Landscape Initiative, by Landscape Chairs [without paper].
- Safety of Artificial Intelligence: A Collaborative Model. John McDermid and Yan Jia.

Session 3: Safe and Value-Aligned Learning in Decision Making

- Choice Set Misspecification in Reward Inference. Rachel Freedman, Rohin Shah and Anca Dragan.
- Safety Augmentation in Decision Trees. Sumanta Dey, Pallab Dasgupta and Briti Gangopadhyay.
- Aligning with Heterogenous Preferences for Kidney Exchange. Rachel Freedman.

Session 4: DNN Testing and Runtime Monitoring

- Is Uncertainty Quantification in Deep Learning Sufficient for Out-of-Distribution Detection? Adrian Schwaiger, Poulami Sinhamahapatra, Jens Gansloser and Karsten Roscher.
- DeepSmartFuzzer: Reward Guided Test Generation For Deep Learning. Samet Demir, Hasan Ferit Eniser and Alper Sen.
- A Comparison of Uncertainty Estimation Approaches in Deep Learning Components for Autonomous Vehicle Applications. Fabio Arnez, Huascar Espinoza, Ansgar Radermacher and François Terrier.
- Increasing the Trustworthiness of Deep Neural Networks via Accuracy Monitoring. Zhihui Shao, Jianyi Yang and Shaolei Ren..

Additionally, AISafety has currently allocated one invited talk, and plans to invite one Keynote speaker and at least two additional invited talks.

Invited Talk

- Nathalie Baracaldo (IBM Research). Security and Privacy Challenges in Federated Learning.

[3] https://www.ai-safety.org

Six posters will be presented in 2-minutes pitches. Five of them will also be part of this volume as poster papers.

Posters

- Robustness as inherent property of datapoints. Catalin-Andrei Ilie, Marius Popescu, and Alin Stefanescu.
- An Efficient Adversarial Attack on Graph Structured Data. Zhengyi Wang and Hang Su [without paper].
- Towards Safe and Reliable Robot Task Planning. Snehasis Banerje.
- Extracting Money from Causal Decision Theorists. Caspar Oesterheld and Vincent Conitzer.
- Ethically Compliant Planning in Moral Autonomous Systems. Justin Svegliato, Samer Nashed and Shlomo Zilberstein.
- Bayesian Model for Trustworthiness Analysis of Deep Learning Classifiers. Andrey Morozov, Emil Valiev, Michael Beyer, Kai Ding, Lydia Gauerhof and Christoph Schorn.

2.2. Second Workshop Day: Landscape

The second-day workshop (AI Safety Landscape) sessions will be organized into by-invitation talks and panels with structured discussions. The by-invitation talks will focus on diverse topics contributing to understand the AI Safety Landscape scientific and technical challenges, industrial and academic opportunities, as well as gaps and pitfalls.

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

3 Acknowledgements

We thank all those who submitted papers to AISafety 2020 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
- Simos Gerasimou, University of York, UK
- Jonas Nilson, NVIDIA, USA
- Brent Harrison, University of Kentucky, USA
- Siddartha Khastgir, University of Warwick, UK
- Carroll Wainwright, Partnership on AI, USA
- Martin Vechev, ETH Zurich, Switzerland
- Sandhya Saisubramanian, University of Massachusetts Amherst, USA
- Alessio R. Lomuscio, Imperial College London, UK
- Rachel Freedman, UC Berkeley, USA
- Brian Tse, Affiliate at University of Oxford, China
- Michael Paulitsch, Intel, Germany
- Rick Salay, University of Toronto, Canada
- Ganesh Pai, NASA Ames Research Center, 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
- Orlando Avila-García, Atos, Spain
- Rob Ashmore, Defence Science and Technology Laboratory, UK
- I-Jeng Wang, Johns Hopkins University, USA
- Chris Allsopp, Frazer-Nash Consultancy, UK
- Francesca Rossi, IBM and University of Padova, Italy
- Ramana Kumar, Google DeepMind, UK
- Javier Ibañez-Guzman, Renault, France
- Jérémie Guiochet, LAAS-CNRS, France
- Raja Chatila, Sorbonne University, France
- Hang Su, Tsinghua University, China
- François Terrier, CEA LIST, France
- Mehrdad Saadatmand, RISE SICS, Sweden
- Alec Banks, Defence Science and Technology Laboratory, UK
- Gopal Sarma, Broad Institute of MIT and Harvard, USA
- Philip Koopman, Carnegie Mellon University, USA
- Roman Nagy, Autonomous Intelligent Driving, 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
- Douglas Lange, Space and Naval Warfare Systems Center Pacific, USA
- Philippa Ryan Conmy, Adelard, UK
- Stefan Kugele, Technical University of Munich, Germany
- Colin Paterson, University of York, UK
- Ashley Llorens, Johns Hopkins University, USA
- Huáscar Espinoza, Commissariat à l´Energie Atomique, France
- John McDermid, University of York, UK
- Xiaowei Huang, University of Liverpool, UK
- Mauricio Castillo-Effen, Lockheed Martin, USA
- Xin 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

Finally, yet importantly, we thank the IJCAI-PRICAI-20 organization for providing an excellent framework for AISafety 2020.