The AAAI-2023 Workshop on Artificial Intelligence Safety (SafeAI-2023)

Gabriel Pedroza[1], Xiaowei Huang[2], Xin Cynthia Chen[3], Andreas Theodorou[4], José Hernández-Orallo[5], Mauricio Castillo-Effen[6], Richard Mallah[7], John McDermid[8]

[1] CEA LIST, Université Paris-Saclay, France
gabriel.pedroza@cea.fr

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

[3] ETH Zurich, Switzerland
chexin@ethz.ch

[4] Umeå University, Sweden
andreas.theodorou@umu.se

[5] Universitat Politècnica de València, Spain
jorallo@upv.es

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

[7] Future of Life Institute, USA
richard@futureoflife.org

[8] University of York, United Kingdom
john.mcdermid@york.ac.uk

Abstract

We summarize the AAAI-2023 Workshop on Artificial Intelligence Safety (SafeAI-2023)[1], held at the 37th AAAI Conference on Artificial Intelligence on February 13-14, 2023 in Washington DC, USA.

Introduction

Safety in Artificial Intelligence (AI) is increasingly becoming a substantial part of AI research, deeply intertwined with the ethical, legal and societal issues associated with AI systems. Even if AI safety is considered a design principle, 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 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 goals with pragmatic solutions, operational with policy issues, and industry with academia, in order to build, evaluate, deploy, operate and maintain AI-based systems that are truly safe.

The AAAI-2023 Workshop on Artificial Intelligence Safety (SafeAI-2023) seeks to explore new ideas in AI safety with a particular focus on addressing the following questions:

- What is the status of existing approaches for ensuring AI and Machine Learning (ML) safety and what are the gaps?
- How can we engineer trustworthy AI software architectures?

[1] Workshop series website: https://safeai.webs.upv.es/

- 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?

These are the main topics of the series of AISafety workshops. They aim to achieve a holistic view of AI and safety engineering, taking ethical and legal issues into account, in order to build trustworthy intelligent autonomous machines. Keeping those topics in our scope, we held the fifth edition of SafeAI, in Washington DC at the 37st AAAI Conference on Artificial Intelligence on February 13-14, 2023. The details of the program are as follows.

Program

The Program Committee (PC) received 48 submissions. Each paper was peer-reviewed by at least two PC members, by following a single-blind reviewing process. The committee decided to accept 17 full papers and 11 poster presentations, resulting in a full-paper acceptance rate of 35% and an overall acceptance rate of 58%.

The SafeAI-2023 program was organized in nine thematic sessions, two (invited) special panel discussions and two (invited) talks. The special panel discussions were given flexibility to structure its program and format.

The thematic sessions followed a highly interactive format. They were structured into short pitches and a group debate 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 sessions and plenary discussions, monitored time, and moderated questions and discussions from the audience.
- *Presenters* gave a 10-minute paper talk and participated in the debate slot. The short presentations are given 5 minutes for each paper.
- *Session Discussants* gave a critical review of the session papers, and participated in the plenary debate.

Presentations and papers were grouped by topic as follows:

Session 1: AI/ML Learning, Explainability, Accuracy and Policy Alignment

- *Active Reward Learning from Multiple Teachers*, Peter Barnett, Rachel Freedman, Justin Svegliato and Stuart Russell
- *REVEALE: Reward Verification and Learning Using Explanations*, Saaduddin Mahmud, Sandhya Saisubramanian and Shlomo Zilberstein
- *A Robust Drift Detection Algorithm with High Accuracy and Low False Positives Rate*, Maxime Fuccellaro, Laurent Simon and Akka Zemmari

Session 2: Short Presentations - Safety Assessment of AI-enabled systems

- *On Evaluating Adversarial Robustness of Chest X-ray Classification*, Salah Ghamizi, Maxime Cordy, Mike Papadakis and Yves Le Traon
- *Multi-timescale Online Monitoring of AI Models*, Fateh Kaakai and Paul-Marie Raffi
- *Capabilities for Better ML Engineering*, Chenyang Yang, Rachel Brower-Sinning, Grace A. Lewis, Christian Kästner and Tongshuang Wu

Session 3: AI/ML for Safety Critical Applications: Assurance Cases and Datasets

- *Transfer Assurance for Machine Learning in Autonomous Systems*, Chiara Picardi, Richard Hawkins, Colin Paterson and Ibrahim Habli
- *Domain-centric ADAS Datasets*, Vaclav Divis, Tobias Schuster and Marek Hruz
- *Towards Developing Safety Assurance Cases for Learning-Enabled Medical Cyber-Physical Systems*, Maryam Bagheri, Josephine Lamp, Xugui Zhou, Lu Feng and Homa Alemzadeh

Session 4 - Short Presentations: ML/DL Robustness: GAM and Attack Detection

- *Evaluation of GAN Architectures for Adversarial Robustness of Convolution Classifier*, Weimin Zhao, Sanaa Alwidian and Qusay Mahmoud
- *Backdoor Attack Detection in Computer Vision by applying Matrix Factorization on the Weights of Deep Networks*, Khondoker Murad Hossain and Tim Oates

Session 5 - AI Safety Assessment: Failure-Cause Analysis, Assurance, Verification

- *A taxonomic system for failure cause analysis of open source AI incidents*, Nikiforos Pittaras and Sean McGregor
- *Towards Safety Assurance of Uncertainty-Aware Reinforcement Learning Agents*, Felippe Schmoeller Roza, Simon Hadwiger, Ingo Thon and Karsten Roscher

- *Formal Verification of Tree Ensembles against Real-World Composite Geometric Perturbations*, Valency Oscar Colaco and Simin Nadjm-Tehrani

Session 6: AI Robustness: Adversarial and Attacks Learning

- *Critically Assessing the State of the Art in CPU-based Local Robustness Verification*, Matthias König, Annelot Bosman, Holger Hoos and Jan van Rijn
- *Towards Understanding How Self-training Tolerates Data Backdoor Poisoning*, Soumyadeep Pal, Ren Wang, Yuguang Yao and Sijia Liu
- *Less is More: Data Pruning for Faster Adversarial Training*, Yize Li, Pu Zhao, Xue Lin, Bhavya Kailkhura and Ryan Goldhahn
- *Personalized Models Resistant to Malicious Attacks for Human-centered Trusted AI*, Teddy Ferdinan and Jan Kocoń

Session 7: AI Robustness: Deep Reinforcement Learning

- *Robustness with Black-Box Adversarial Attack using Reinforcement Learning*, Soumyendu Sarkar, Ashwin Ramesh Babu, Sajad Mousavi, Vineet Gundecha, Sahand Ghorbanpour, Alexander Shmakov, Ricardo Luna Gutierrez, Antonio Guillen and Avisek Naug
- *White-Box Adversarial Policies in Deep Reinforcement Learning*, Stephen Casper, Dylan Hadfield-Menell and Gabriel Kreiman
- *Bab: A novel algorithm for training clean model based on poisoned data*, Chen Chen, Hong Haibo, Xie Mande, Jun Shao and Tao Xiang
- *Safe Reinforcement Learning through Phasic Safety-Oriented Policy Optimization*, Sumanta Dey, Pallb Dasgupta and Soumyajit Dey

Session 8 - Short Presentations: OoD Detection and Uncertainty for ML/DL Safety

- *Out-of-Distribution Detection Using Deep Neural Network Latent Space Uncertainty*, Fabio Arnez, Ansgar Radermacher and François Terrier
- *Efficient and Effective Uncertainty Quantification in Gradient Boosting via Cyclical Gradient MCMC*, Tian Tan, Carlos Huertas and Qi Zhao
- *Safety Assurance with Ensemble-based Uncertainty Estimation and overlapping alternative Predictions in Reinforcement Learning*, Dirk Eilers, Simon Burton, Felippe Schmoeller da Roza and Karsten Roscher

Session 9 - Short Presentations: Methods and Techniques for AI/ML Safety Assessment

- *Towards a holistic approach for AI trustworthiness assessment based upon aids for multi-criteria aggregation*, Juliette Mattioli, Henri Sohier, Agnes Delaborde, Gabriel Pedroza, Kahina Amokrane-Ferka, Afef Awadid, Zakaria Chihani and Souhaiel Khalfaoui
- *A Framework Quantifying Trustworthiness of Supervised Machine and Deep Learning Models*, Alberto Huertas Celdran, Jan Kreischer, Melike Demirci, Joel Leupp, Pedro Miguel Sanchez, Muriel Franco, Gerome Bovet, Gregorio Martinez Perez and Burkhard Stiller
- *Standardizing the Probabilistic Sources of Uncertainty for the sake of Safety Deep Learning*, Axel Brando, Isabel Serra, Enrico Mezzetti, Francisco J. Cazorla and Jaume Abella

SafeAI was pleased to have several additional inspirational researchers as invited speakers:

Distinctive Panel Discussion

- Stuart Russell (UC Berkeley), Song-Chun Zhu (PKU, THU, and BIGAI), Mark Nitzberg (UC Berkeley), *On Artificial General Intelligence and AI Alignment*

Special Panel Discussion

- Fateh Kaakai (Thales, IRT SystemX), Souhaiel Khalfaoui (Valeo), Augustin Lemesle (CEA), *Towards trustworthiness of AI-enabled systems: the Confiance.ai Programme*

Invited Talks

- Martin Rothfelder (Siemens AG), *Digitization and automation for driverless regional trains – The safe.trAIn research project*
- Vincent Conitzer (Carnegie Mellon U., U. of Oxford), *Foundations of Cooperative AI*

Acknowledgements

We thank all researchers who submitted papers to SafeAI-2023 and congratulate the authors whose papers were selected for inclusion into the workshop program and proceedings.

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

- Huáscar Espinoza, KDT JU, Belgium
- Stuart Russell, UC Berkeley, USA
- Raja Chatila, Sorbonne University, France
- Francesca Rossi, IBM and University of Padova, USA
- Roman V. Yampolskiy, University of Louisville, USA
- Gereon Weiss, Fraunhofer IKS, Germany
- Roman Nagy, Argo AI, Germany
- Nathalie Baracaldo, IBM Research, USA

- Chokri Mraidha, CEA LIST, France
- Brent Harrison, University of Kentucky, USA
- Toshihiro Nakae, DENSO Corporation, Japan
- John Favaro, Trust-IT, Italy
- Agnes Delaborde, LNE, France
- Jonas Nilsson, NVIDIA, USA
- Leon Kester, TNO, The Netherlands
- Michael Paulitsch, Intel, Germany
- Philippa Ryan Conmy, University of York, UK
- Stefan Kugele, Technische Hochschule Ingolstadt, Germany
- Javier Ibañez-Guzman, Renault, France
- Mehrdad Saadatmand, RISE SICS, Sweden
- Alessio R. Lomuscio, Imperial College London, UK
- Jérémie Guiochet, LAAS-CNRS, France
- Sandhya Saisubramanian, University of Massachusetts Amherst, USA
- Mario Gleirscher, University of Bremen, Germany
- Chris Allsopp, Origami Labs, UK
- Vahid Behzadan, University of New Haven, USA
- Simos Gerasimou, University of York, UK
- Feng Liu, Huawei Munich Research Center, Germany
- Juliette Mattioli, Thales, France
- Brian Tse, Affiliate at University of Oxford, China
- Colin Paterson, University of York, UK
- Peter Flach, University of Bristol, UK
- Simon Fuerst, BMW Group, Germany
- Emmanuelle Escorihuela, Airbus, France
- Roel Dobbe, TU Delft, The Netherlands
- Andrea Orlandini, ISTC-CNR, Italy
- Ke Pei, Huawei, China
- Mohamed Ibn Khedher, IRT SystemX, France
- Ganesh Pai, NASA Ames Research Center, USA
- Davide Bacciu, Università di Pisa, Italy
- Rasmus Adler, Fraunhofer IESE, Germany
- Danilo Vasconcellos Vargas, Kyushu University, Japan
- Vahid Hashemi, Audi, Germany
- Umut Durak, German Aerospace Center (DLR), Germany
- Morayo Adedjouma, CEA LIST, France
- John Burden, University of Cambridge, UK
- Luciano Cavalcante Siebert, TU Delft, The Netherlands
- Timo Sämann, Valeo, Germany
- Jan Reich, Fraunhofer IESE, Germany
- Mandar Pitale, NVIDIA, USA
- Nikolaos Matragkas, CEA LIST, France
- Bowei Xi, University of Purdue, USA
- Fateh Kaakai, Thales, France

We thank Stuart Russell, Song-Chun Zhu, Mark Nitzberg, Martin Rothfelder, Vincent Conitzer, Fateh Kaakai, Souhaiel Khalfaoui, and Augustin Lemesle for their inspiring talks.

Finally we thank the AAAI-2023 organization for providing an excellent framework for SafeAI-2023.