The AAAI-22 Workshop on Artificial Intelligence Safety (SafeAI 2022)

Gabriel Pedroza[1], José Hernández-Orallo[2], Xin Cynthia Chen[3], Xiaowei Huang[4], Huáscar Espinoza[5], Mauricio Castillo-Effen[6], John McDermid[7], Richard Mallah[8], Seán S. ÓhÉigeartaigh[9]

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

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

[3] University of Hong Kong, China
cyn0531@connect.hku.hk

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

[5] ECSEL JU, Belgium
Huascar.Espinoza@ecsel.europa.eu

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

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

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

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

Abstract

We summarize the AAAI-22 Workshop on Artificial Intelligence Safety (SafeAI 2022)[1], virtually held at the Thirty-Sixth AAAI Conference on Artificial Intelligence on February 28.

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-22 Workshop on Artificial Intelligence Safety (SafeAI 2022) seeks to explore new ideas in AI

[1] Workshop series website: http://safeaiw.org/

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?
- 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 SafeAI 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. 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), and the second edition was held in February 7, 2020 in New York City (USA) also as part of AAAI. This fourth edition was held online (because of the COVID-19 situation) at the Thirty-Sixth AAAI Conference on Artificial Intelligence on February 28, virtually.

Program

The Program Committee (PC) received 53 submissions. Each paper was peer-reviewed by at least two PC members, by following a single-blind reviewing process. The committee decided to accept 18 full papers, 3 talks and 12 posters, resulting in a full-paper acceptance rate of 34.0% and an overall acceptance rate of 62.3%.

The SafeAI 2022 program was organized in five thematic sessions, two keynotes and three (invited) talks.

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 minutes paper talk and participated in the debate slot.
- *Session Discussants* gave a critical review of the session papers, and participated in the plenary debate.

Papers were grouped by topic as follows:

Session 1: Bias, Fairness and Value Alignment

- The Problem of Behaviour and Preference Manipulation in AI Systems, Hal Ashton and Matija Franklin.
- IFBiD: Inference-Free Bias Detection, Ignacio Serna, Daniel DeAlcala, Aythami Morales Moreno, Julian Fierrez and Javier Ortega-Garcia.
- Blackbox Post-Processing for Multiclass Fairness, Preston Putzel and Scott Lee.

Session 2: Interpretability and Accountability

- A Gray Box Model for Characterizing Driver Behavior, Soyeon Jung, Ransalu Senanayake and Mykel Kochenderfer.
- Defining and Identifying the Legal Culpability of Side Effects using Causal Graphs, Hal Ashton.

Session 3: Robustness and Uncertainty

- Efficient Adversarial Sequence Generation for RNN with Symbolic Weighted Finite Automata, Mingjun Ma, Dehui Du, Yuanhao Liu, Yanyun Wang and Yiyang Li.
- A Study on Mitigating Hard Boundaries of Decision-Tree-based Uncertainty Estimates for AI Models, Pascal Gerber, Lisa Jöckel and Michael Kläs.
- Quantifying the Importance of Latent Features in Neural Networks, Amany Alshareef, Nicolas Berthier, Sven Schewe and Xiaowei Huang.
- Maximum Likelihood Uncertainty Estimation: Robustness to Outliers, Deebul Nair, Nico Hochgeschwender and Miguel Olivares-Mendez.

Session 4: Safe Reinforcement Learning

- Reinforcement Learning With Imperfect Safety Constraints, Jin Woo Ro, Gerald Lüttgen and Diedrich Wolter.
- Do Androids Dream of Electric Fences? Safety-Aware Reinforcement Learning with Latent Shielding, Peter He, Borja Leon and Francesco Belardinelli.
- HiSaRL: A Hierarchical Framework for Safe Reinforcement Learning, Zikang Xiong, Ishika Agarwal and Suresh Jagannathan.
- A Game-Theoretic Perspective on Risk-Sensitive Reinforcement Learning, Mathieu Godbout, Maxime Heuillet, Sharath Chandra Raparthy, Rupali Bhati and Audrey Durand.

Session 5: AI Testing and Assessment

- Beyond Test Accuracy: The Effects of Model Compression on CNNs, Adrian Schwaiger, Kristian Schwienbacher and Karsten Roscher.
- Differential Assessment of Black-Box AI Agents, Rashmeet Kaur Nayyar, Pulkit Verma and Siddharth Srivastava. (Note that this paper is out of the proceedings of SafeAI2022.)
- Using Adaptive Stress Testing to Identify Paths to Ethical Dilemmas in Autonomous Systems, Ann-Katrin Reuel, Mark Koren, Anthony Corso and Mykel J. Kochenderfer.

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

Keynote

- Matthew Dwyer (University of Virginia), Distribution-aware Test Adequacy for Neural Networks
- Ganesh Pai (KBR / NASA Ames Research Center), Towards Certification of Machine Learning in Aeronautical Applications

Invited Talks

- Shiri Dori-Hacohen (University of Connecticut), Quantifying Misalignment Between Agents
- Roel Dobbe (TU Delft), A System Safety Perspective for Developing and Governing Artificial Intelligence
- Bonnie W. Johnson (Naval Postgraduate School), Safety in AI-Enabled Warfare Decision Aids

Posters

Posters were presented with 2-minute pitches. Most posters have also been included as short papers within this volume.

- Near-Term AI as an Existential Risk Factor, Ben Bucknall and Shiri Dori-Hacohen. (Note that this paper is out of the proceedings of SafeAI2022.)
- The Dilemma Between Data Transformations and Adversarial Robustness for Time Series Application Systems, Sheila Alemany and Niki Pissinou.
- Interpretable Local Tree Surrogate Policies, John Mern, Sidhart Krishnan, Anil Yildiz, Kyle Hatch and Mykel J. Kochenderfer.
- Oases of Cooperation: An Empirical Evaluation of Reinforcement Learning in the Iterated Prisoner's Dilemma, Peter Barnett and John Burden.
- Leveraging Multi-task Learning for Umambiguous and Flexible Deep Neural Network Watermarking, Fangqi Li, Lei Yang, Shilin Wang and Alan Wee-Chung Liew.
- Human-in-the-loop Learning for Safe Exploration through Anomaly Prediction and Intervention, Prajit T Rajendran, Huascar Espinoza, Agnes Delaborde and Chokri Mraidha.
- Safety Aware Reinforcement Learning by Identifying Comprehensible Constraints in Expert Demonstrations, Leopold Müller, Lars Böcking and Michael Färber.
- Combining Data-Driven and Knowledge-Based AI Paradigms for Engineering AI-Based Safety-Critical Systems, Juliette Mattioli, Gabriel Pedroza, Souhaiel Khalfaoui and Bertrand Leroy.
- Is it all a cluster game? – Exploring Out-of-Distribution Detection based on Clustering in the Embedding Space, Poulami Sinhamahapatra, Rajat Koner, Karsten Roscher and Stephan Günnemann.
- A Practical Overview of Safety Concerns and Mitigation Methods for Visual Deep Learning Algorithms, Saeed Bakhshi Germi and Esa Rahtu.
- Comparing Vision Transformers and Convolutional Nets for Safety Critical Systems, Michal Filipiuk and Vasu Singh.
- A Framework to Argue Quantitative Safety Targets in Assurance Cases for AI/ML Components Combining Design and Runtime Safety Measures, Michael Klaes, Lisa Jöckel, Rasmus Adler and Jan Reich. (Note that this paper is out of the proceedings of SafeAI2022.)

Special Sessions

- EnnCore: End-to-End Conceptual Guarding of Neural Architectures, Edoardo Manino, Danilo Carvalho, Yi Dong, Julia Rozanova, Xidan Song, Andre Freitas, Gavin Brown, Mikel Luján, Xiaowei Huang, and Lucas Cordeiro.
- The wall of safety for AI: approaches in the Confiance.ai program, Bertrand Braunschweig, François Terrier, and Rodolphe Gelin.

Acknowledgements

We thank all researchers who submitted papers to SafeAI 2022 and congratulate the authors whose papers and posters 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:

- 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 ESK, 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, Adelard, 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 York, UK
- Chris Allsopp, Frazer-Nash Consultancy, 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
- Gabriel Pedroza, CEA LIST, France
- Huáscar Espinoza, ECSEL JU, Belgium
- Xin Cynthia Chen, University of Hong Kong, China
- Xiaowei Huang, University of Liverpool, UK
- José Hernández-Orallo, Universitat Politècnica de València, Spain
- Mauricio Castillo-Effen, Lockheed Martin, USA
- John McDermid, University of York, UK
- Richard Mallah, Future of Life Institute, USA
- Seán Ó hÉigeartaigh, University of Cambridge, UK

We thank Matthew Dwyer, Ganesh Pai, Shiri Dori-Hacohen, Roel Dobbe, and Bonnie W. Johnson for their inspiring talks.

We also want to thank Jessica Schmidt, T R Prajit, Sarra Bouchkati, Jordan Hamilton, Fabio Arnez, and Megha Quamara for their help in reviewing the papers.

Finally we thank the AAAI-22 organization for providing an excellent framework for SafeAI 2022.