The IJCAI-ECAI-22 Workshop on Artificial Intelligence Safety (AISafety2022)

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

[1] CEA LIST, France
gabriel.pedroza@cea.fr

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

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

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

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

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

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

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

Abstract

We summarize the IJCAI-ECAI-22 Workshop on Artificial Intelligence Safety (AISafety 2022)[1], held at the 31st International Joint Conference on Artificial Intelligence and the 25th European Conference on Artificial Intelligence (IJCAI-ECAI-22) on July 24-25, 2022 in Vienna, Austria.

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

[1] Workshop series website: https://www.aisafetyw.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 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. The first edition of AISafety was held in August 10-12, 2019, in Macao (China) as part of the 28th International Joint Conference on Artificial Intelligence (IJCAI-19), and the second edition was held in January 7-8, 2021 virtually also as part of IJCAI. This fourth edition was held in Vienna at the 31st International Joint Conference on Artificial Intelligence (IJCAI-ECAI-22) on July 24-25th.

Program

The Program Committee (PC) received 26 submissions. Each paper was peer-reviewed by at least two PC members, by following a single-blind reviewing process. The committee decided to accept 13 full papers and 6 short presentations, resulting in a full-paper acceptance rate of 50% and an overall acceptance rate of 73%.

The AISafety 2022 program was organized in six thematic sessions, one (invited) special session, two keynote and four (invited) talks. The special session was 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 Ethics: Fairness, Bias, and Accountability

- Let it RAIN for Social Good, Mattias Brännström, Andreas Theodorou and Virginia Dignum.
- Accountability and Responsibility of Artificial Intelligence Decision-making Models in Indian Policy Landscape, Palak Malhotra and Amita Misra.
- Assessing Demographic Bias Transfer from Dataset to Model: A Case Study in Facial Expression Recognition, Iris Dominguez-Catena, Daniel Paternain and Mikel Galar.

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

- A Hierarchical HAZOP-Like Safety Analysis for Learning-Enabled Systems, Yi Qi, Philippa Ryan Conmy, Wei Huang, Xingyu Zhao and Xiaowei Huang.
- Increasingly Autonomous CPS: Taming Emerging Behaviors from an Architectural Perspective, Jerome Hugues and Daniela Cancila.
- CAISAR: A platform for Characterizing Artificial Intelligence Safety and Robustness, Julien Girard-Satabin, Michele Alberti, François Bobot, Zakaria Chihani and Augustin Lemesle.

Session 3: Machine learning for safety-critical AI

- Revisiting the Evaluation of Deep Neural Networks for Pedestrian Detection, Patrick Feifel, Benedikt Franke, Arne Raulf, Friedhelm Schwenker, Frank Bonarens and Frank Köster.
- Improvement of Rejection for AI Safety through Loss-Based Monitoring, Daniel Scholz, Florian Hauer, Klaus Knobloch and Christian Mayr.

Special Session : TAILOR - Towards Trustworthy AI

- Foundations of Trustworthy AI*, Francesca Pratesi.
- Panel on Trustworthy AI*, Fosca Giannotti, Pilipp Slusallek, Giuseppe De Giacomo, Hector Geffner, Holger Hoos.

**Presentations without papers.*

Session 4: Short Presentations - ML Robustness, Criticality and Uncertainty

- Utilizing Class Separation Distance for the Evaluation of Corruption Robustness of Machine Learning Classifiers, Georg Siedel, Silvia Vock, Andrey Morozov and Stefan Voß.
- Safety-aware Active Learning with Perceptual Ambiguity and Criticality Assessment, Prajit T Rajendran, Guillaume Ollier, Huascar Espinoza, Morayo Adedjouma, Agnes Delaborde and Chokri Mraidha.
- Understanding Adversarial Examples Through Deep Neural Network's Classification Boundary and Uncertainty Regions, Juan Shu, Bowei Xi and Charles Kamhoua.

- **Session 5: AI Robustness, Generative models and Adversarial learning**

- Leveraging generative models to characterize the failure conditions of image classifiers, Adrien Le Coz, Stéphane Herbin and Faouzi Adjed.
- Feasibility of Inconspicuous GAN-generated Adversarial Patches against Object Detection, Svetlana Pavlitskaya, Bianca-Marina Codău and J. Marius Zöllner.
- Privacy Safe Representation Learning via Frequency Filtering Encoder, Jonghu Jeong, Minyong Cho, Philipp Benz, Jinwoo Hwang, Jeewook Kim, Seungkwan Lee and Tae-hoon Kim.
- Benchmarking and deeper analysis of adversarial patch attack on object detectors, Pol Labarbarie, Adrien Chan Hon Tong, Stéphane Herbin and Milad Leyli-Abadi.

Session 6: AI Accuracy, Diversity, Causality and Optimization

- The impact of averaging logits over probabilities on ensembles of neural networks, Cedrique Rovile Njieutcheu Tassi, Jakob Gawlikowski, Auliya Unnisa Fitri and Rudolph Triebel.
- Exploring Diversity in Neural Architectures for Safety, Michał Filipiuk and Vasu Singh.
- Constrained Policy Optimization for Controlled Contextual Bandit Exploration, Mohammad Kachuee and Sungjin Lee.
- A causal perspective on AI deception in games, Francis Rhys Ward, Francesco Belardinelli and Francesca Toni.

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

Keynotes

- Gary Marcus, Towards a Proper Foundation for Robust Artificial Intelligence
- Thomas A. Henzinger, Formal Methods meet Neural Networks: A Selection

Invited Talks

- Elizabeth Adams, Leadership of Responsible AI – Representation Matters
- Luis Aranda, Enabling AI governance: OECD's work on moving from Principles to practice
- Simos Gerasimou, SESAME: Secure and Safe AI-Enabled Robotics Systems
- Zakaria Chihani, A selected view of AI trustworthiness methods: How far can we go?

Acknowledgements

We thank all researchers who submitted papers to AISafety 2022 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:

- Simos Gerasimou, University of York, UK
- Jonas Nilson, NVIDIA, USA
- Morayo Adedjouma, CEA LIST, France
- Brent Harrison, University of Kentucky, USA
- Alessio R. Lomuscio, Imperial College London, UK
- Brian Tse, Affiliate at University of Oxford, China
- Michael Paulitsch, Intel, Germany
- Ganesh Pai, NASA Ames Research Center, USA
- Rob Alexander, University of York, UK
- Vahid Behzadan, University of New Haven, USA
- Chokri Mraidha, CEA LIST, France
- Ke Pei, Huawei, China
- Orlando Avila-García, Arquimea Research Center, Spain
- I-Jeng Wang, Johns Hopkins University, USA
- Chris Allsopp, Frazer-Nash Consultancy, UK
- Andrea Orlandini, ISTC-CNR, Italy
- Agnes Delaborde, LNE, France
- Rasmus Adler, Fraunhofer IESE, Germany
- Roel Dobbe, TU Delft, The Netherlands
- Vahid Hashemi, Audi, Germany
- Juliette Mattioli, Thales, France
- Bonnie W. Johnson, Naval Postgraduate School, USA
- Roman V. Yampolskiy, University of Louisville, USA
- Jan Reich, Fraunhofer IESE, Germany
- Fateh Kaakai, Thales, France
- Francesca Rossi, IBM and University of Padova, USA
- Javier Ibañez-Guzman, Renault, France
- Jérémie Guiochet, LAAS-CNRS, France
- Raja Chatila, Sorbonne University, France

- François Terrier, CEA LIST, France
- Mehrdad Saadatmand, RISE Research Institutes of Sweden, Sweden
- Alec Banks, Defence Science and Technology Laboratory, UK
- Roman Nagy, Argo AI, Germany
- Nathalie Baracaldo, IBM Research, USA
- Toshihiro Nakae, DENSO Corporation, Japan
- Gereon Weiss, Fraunhofer ESK, Germany
- Philippa Ryan Conmy, Adelard, UK
- Stefan Kugele, Technische Hochschule Ingolstadt, Germany
- Colin Paterson, University of York, UK
- Davide Bacciu, Università di Pisa, Italy
- Timo Sämann, Valeo, Germany
- Sylvie Putot, Ecole Polytechnique, France
- John Burden, University of Cambridge, UK
- Sandeep Neema, DARPA, USA
- Fredrik Heintz, Linköping University, Sweden
- Simon Fürst, BMW Group, Germany
- Mario Gleirscher, University of Bremen, Germany
- Mandar Pitale, NVIDIA, USA
- Leon Kester, TNO, The Netherlands
- Gabriel Pedroza, CEA LIST, France
- Huáscar Espinoza, KDT JU, Belgium
- Xiaowei Huang, University of Liverpool, UK
- José Hernández-Orallo, Universitat Politècnica de València, Spain
- Mauricio Castillo-Effen, Lockheed Martin, USA
- Xin Cynthia Chen, University of Hong Kong, China
- Richard Mallah, Future of Life Institute, USA
- John McDermid, University of York, United Kingdom

We thank Gary Marcus, Thomas A. Henzinger, Elizabeth Adams, Luis Aranda, Simos Gerasimou, and Zakaria Chihani for their inspiring talks.

Finally we thank the IJCAI-ECAI-22 organization for providing an excellent framework for AISafety 2022.