The IJCAI-21 Workshop on Artificial Intelligence Safety (AISafety2021)

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

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

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

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

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

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

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

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

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

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

Abstract

We summarize the IJCAI-21 Workshop on Artificial Intelligence Safety (AISafety 2021)[1], virtually held at the 30th International Joint Conference on Artificial Intelligence (IJCAI-21) on August 19-20, 2021.

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,

[1] Workshop series website: https://www.aisafetyw.org/

deploy, operate and maintain AI-based systems that are truly safe.

The IJCAI-21 Workshop on Artificial Intelligence Safety (AISafety 2021) 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?
- 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 third edition was held online (because of the COVID-19 situation) at the 30th International Joint Conference on Artificial Intelligence (IJCAI-21) on August 19-20, virtually.

Program

The Program Committee (PC) received 25 submissions. Each paper was peer-reviewed by at least two PC members, by following a single-blind reviewing process. The committee decided to accept 11 full papers and 7 posters, resulting in a full-paper acceptance rate of 44% and an overall acceptance rate of 72%.

The AISafety 2021 program was organized in four thematic sessions, two keynote and two (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 minute 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: Trustworthiness of Knowledge-Based AI

- Applying Strategic Reasoning for Accountability Ascription in Multiagent Teams, Vahid Yazdanpanah, Sebastian Stein, Enrico Gerding and Nicholas R. Jennings.
- Impossibility of Unambiguous Communication as a Source of Failure in AI Systems, William Howe and Roman Yampolskiy.

Session 2: Robustness of Machine Learning Approaches

- Assessing the Reliability of Deep Learning Classifiers Through Robustness Evaluation and Operational Profiles, Xingyu Zhao, Wei Huang, Alec Banks, Victoria Cox, David Flynn, Sven Schewe and Xiaowei Huang.
- Towards Robust Perception Using Topological Invariants, Romie Banerjee, Feng Liu and Pei Ke.
- Measuring Ensemble Diversity and Its Effects on Model Robustness, Lena Heidemann, Adrian Schwaiger and Karsten Roscher.

Session 3: Perception and Adversarial Attacks

- Deep Neural Network Loses Attention to Adversarial Images, Shashank Kotyan and Danilo Vasconcellos Vargas.
- An Adversarial Attacker for Neural Networks in Regression Problems, Kavya Gupta, Jean-Christophe Pesquet, Beatrice Pesquet-Popescu, Fateh Kaakai and Fragkiskos Malliaros.
- Coyote: A Dataset of Challenging Scenarios in Visual Perception for Autonomous Vehicles, Suruchi Gupta, Ihsan Ullah and Michael Madden.

Session 4: Qualification / Certification of AI-Based Systems

- Towards a Safety Case for Hardware Fault Tolerance in Convolutional Neural Networks Using Activation Range Supervision, Florian Geissler, Syed Qutub, Sayanta Roychowdhury, Ali Asgari, Yang Peng, Akash Dhamasia, Ralf Graefe, Karthik Pattabiraman and Michael Paulitsch.
- Artificial Intelligence for Future Skies: On-going Standardization Activities to Build the Next

Certification/Approval Framework for Airborne and Ground Aeronautic Products, Christophe Gabreau, Béatrice Pesquet-Popescu, Fateh Kaakai and Baptiste Lefevre.
- Using Complementary Risk Acceptance Criteria to Structure Assurance Cases for Safety-Critical AI Components, Michael Klaes, Rasmus Adler, Lisa Jöckel, Janek Groß and Jan Reich.

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

Keynotes

- Emily Dinan (Facebook AI Research, USA), Safety for E2E Conversational AI
- Simon Burton (Fraunhofer IKS, Germany), Safety, Complexity, AI and Automated Driving - Holistic Perspectives on Safety Assurance

Invited Talks
- The Anh Han (Teesside University, UK), Modelling and Regulating Safety Compliance: Game Theory Lessons from AI Development Races Analyses
- Umut Durak (German Aerospace Center - DLR, Germany), Simulation Qualification for Safety Critical AI-Based Systems

Posters were presented with 3-minute pitches. Most posters have also been included as short papers within this volume.

Posters

- Uncontrollability of Artificial Intelligence, Roman Yampolskiy.
- Domain Shifts in Reinforcement Learning: Identifying Disturbances in Environments, Tom Haider, Felippe Schmoeller Roza, Dirk Eilers, Karsten Roscher and Stephan Günnemann.
- Chess as a Testing Grounds for the Oracle Approach to AI Safety, James Miller, Roman Yampolskiy, Olle Häggström and Stuart Armstrong.
- Socio-technical co-Design for Accountable Autonomous Software, Ayan Banerjee, Imane Lamrani, Katina Michael, Diana Bowman and Sandeep Gupta.
- Epistemic Defenses against Scientific and Empirical Adversarial AI Attacks, Nadisha-Marie Aliman and Leon Kester.
- On the Differences between Human and Machine Intelligence, Roman Yampolskiy.
- A Mixed Integer Programming Approach for Verifying Properties of Binarized Neural Networks, Christopher Lazarus and Mykel Kochenderfer.

Acknowledgements

We thank all researchers who submitted papers to AISafety 2021 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
- Emmanuel Arbaretier, Apsys-Airbus, France
- Ann Nowé, Vrije Universiteit Brussel, Belgium
- Simos Gerasimou, University of York, UK
- Gereon Weiss, Fraunhofer ESK, Germany
- 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
- Hélène Waeselynck, CNRS LAAS, France
- Rob Alexander, University of York, UK
- Vahid Behzadan, Kansas State University, USA
- Chokri Mraidha, CEA LIST, France
- Ke Pei, Huawei, China
- Orlando Avila-García, Arquimea Research Center, Spain
- Rob Ashmore, Defence Science and Technology Laboratory, UK
- I-Jeng Wang, Johns Hopkins University, USA
- Chris Allsopp, Frazer-Nash Consultancy, UK
- Andrea Orlandini, ISTC-CNR, Italy
- Rasmus Adler, Fraunhofer IESE, Germany
- Roel Dobbe, TU Delft, The Netherlands
- Vahid Hashemi, Audi, Germany
- Feng Liu, Huawei Munich Research Center, Germany
- Yogananda Jeppu, Honeywell Technology Solutions, India
- Francesca Rossi, IBM and University of Padova, USA
- Ramana Kumar, Google DeepMind, UK
- 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
- Gopal Sarma, Broad Institute of MIT and Harvard, USA
- Roman Nagy, Argo AI, Germany
- Nathalie Baracaldo, IBM Research, USA
- Toshihiro Nakae, DENSO Corporation, Japan
- Richard Cheng, California Institute of Technology, USA
- 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, Technische Hochschule Ingolstadt, Germany
- Colin Paterson, University of York, UK
- Javier Garcia, Universidad Carlos III de Madrid, Spain
- Davide Bacciu, Università di Pisa, Italy
- Timo Sämann, Valeo, Germany
- Vincent Aravantinos, Argo AI, Germany
- Mohamed Ibn Khedher, IRT SystemX, France
- Umut Durak, German Aerospace Center (DLR), Germany
- Huáscar Espinoza, ECSEL JU
- Seán Ó hÉigeartaigh, University of Cambridge, UK
- 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
- Gabriel Pedroza, CEA LIST, France

As well as the additional reviewers:
- Fabio Arnez, CEA LIST, France
- Emmanouil Seferis, Audi, Germany
- Joris Guerin, LAAS CNRS, France

We thank Emily Dinan, Simon Burton, The Anh Han and Umut Durak for their inspiring talks.

We would like to specially thank our sponsor, Partnership on AI, which funded the Best Paper Award.

Finally we thank the IJCAI-21 organization for providing an excellent framework for AISafety 2021.