The AAAI-21 Workshop on Artificial Intelligence Safety (SafeAI 2021)

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

[1] CEA LIST, Gif-sur-Yvette, France
huascar.espinoza@cea.fr

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

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

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

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

[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-21 Workshop on Artificial Intelligence Safety (SafeAI 2021)[1], virtually held at the Thirty-Fifth AAAI Conference on Artificial Intelligence on February 8.

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-21 Workshop on Artificial Intelligence Safety (SafeAI 2021) seeks to explore new ideas in AI safety with a particular focus on addressing the following questions:

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

- 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 third edition was held online (because of the COVID-19 situation) at the Thirty-Fifth AAAI Conference on Artificial Intelligence on February 8, virtually.

Program

The Program Committee (PC) received 44 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, 1 talks and 11 posters, resulting in a full-paper acceptance rate of 29.5% and an overall acceptance rate of 56.8%.

The SafeAI 2021 program was organized in four thematic sessions, one 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: Dynamic Safety and Anomaly Assessment

- Feature Space Singularity for Out-of-Distribution Detection, Haiwen Huang, Zhihan Li, Lulu Wang, Sishuo Chen, Xinyu Zhou and Bin Dong.
- An Evaluation of "Crash Prediction Networks" (CPN) for Autonomous Driving Scenarios in CARLA Simulator, Saasha Nair, Sina Shafaei, Daniel Auge and Alois Knoll.
- From Black-box to White-box: Examining Confidence Calibration under different Conditions, Franziska Schwaiger, Maximilian Henne, Fabian Küppers, Felippe Schmoeller Roza, Karsten Roscher and Anselm Haselhoff.

Session 2: Safety Considerations for the Assurance of AI-based Systems

- The Utility of Neural Network Test Coverage Measures, Rob Ashmore and Alec Banks.
- Safety Properties of Inductive Logic Programming, Gavin Leech, Nandi Schoots and Joar Skalse.
- A Hybrid-AI Approach for Competence Assessment of Automated Driving functions, Jan-Pieter Paardekooper, Mauro Comi, Corrado Grappiolo, Ron Snijders, Willeke van Vught and Rutger Beekelaar.

Session 3: Adversarial Machine Learning and Trustworthiness

- Adversarial Robustness for Face Recognition: How to Introduce Ensemble Diversity among Feature Extractors?, Takuma Amada, Kazuya Kakizaki, Seng Pei Liew, Toshinori Araki, Joseph Keshet and Jun Furukawa.
- Multi-Modal Generative Adversarial Networks Make Realistic and Diverse but Untrustworthy Predictions When Applied to Ill-posed Problems, John Hyatt and Michael Lee.
- DeepFakesON-Phys: DeepFakes Detection based on Heart Rate Estimation, Javier Hernandez-Ortega, Ruben Tolosana, Julian Fierrez and Aythami Morales.

Session 4: Safe Autonomous Agents

- What Criminal and Civil Law Tells Us About Safe RL Techniques to Generate Law-abiding Behaviour, Hal Ashton.

- Performance of Bounded-Rational Agents With the Ability to Self-Modify, Jakub Tětek, Marek Sklenka and Tomáš Gavenčiak.
- Deep CPT-RL: Imparting Human-Like Risk Sensitivity to Artificial Agents, Jared Markowitz, Marie Chau and I-Jeng Wang.
- Challenges for Using Impact Regularizers to Avoid Negative Side Effects, David Lindner, Kyle Matoba and Alexander Meulemans.

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

Keynote

- Not finalized at the date of publishing

Invited Talks

- Juliette Mattioli (Thales, France) and Rodolphe Gelin (Renault, France). Methods and Tools for Trusted AI: an Urgent Challenge for Industry
- Sandhya Saisubramanian (University of Massachusetts Amherst, USA), Challenges and Directions in Avoiding Negative Side Effects

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

Posters

- Towards an Ontological Framework for Environmental Survey Hazard Analysis of Autonomous Systems, Christopher Harper and Praminda Caleb-Solly.
- Overestimation learning with guarantees, Adrien Gauffriau, François Malgouyres and Mélanie Ducoffe.
- On the Use of Available Testing Methods for Verification & Validation of AI-based Software and Systems, Franz Wotawa.
- Runtime Decision Making Under Uncertainty in Autonomous Vehicles, Vibhu Gautam, Youcef Gheraibia, Rob Alexander and Richard Hawkins.
- Negative Side Effects and AI Agent Indicators: Experiments in SafeLife, John Burden, Jose Hernandez-Orallo and Sean O'Heigeartaigh.
- Time for AI (Ethics) Maturity Model Is Now, Ville Vakkuri, Marianna Jantunen, Erika Halme, Kai-Kristian Kemell, Anh Nguyen-Duc, Tommi Mikkonen and Pekka Abrahamsson.
- AI-Blueprint for Deep Neural Networks, Ernest Wozniak, Henrik Putzer and Carmen Carlan.
- Neural Criticality: Validation of Convolutional Neural Networks, Vaclav Divis and Marek Hruz.
- Adversarial Attacks for Tabular Data: Application to Fraud Detection and Imbalanced Data, Francesco Cartella, Orlando Anunciacao, Yuki Funabiki, Daisuke Yamaguchi, Toru Akishita and Olivier Elshocht.
- Correct-by-Construction Multi-Label Classification Networks, Eleonora Giunchiglia and Thomas Lukasiewicz.
- Classification Confidence Scores with Point-wise Guarantees, Nivasini Ananthakrishnan, Shai Ben-David and Tosca Lechner.

Acknowledgements

We thank all researchers who submitted papers to SafeAI 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
- Francesca Rossi, IBM and University of Padova, Italy
- Raja Chatila, Sorbonne University, France
- Roman V. Yampolskiy, University of Louisville, USA
- Gereon Weiss, Fraunhofer ESK, Germany
- Mark Nitzberg, Center for Human-Compatible AI, USA
- Roman Nagy, Autonomous Intelligent Driving GmbH, Germany
- François Terrier, CEA LIST, France
- Hélène Waeselynck, LAAS-CNRS, France
- Siddartha Khastgir, University of Warwick, UK
- Orlando Avila-García, Atos, Spain
- Nathalie Baracaldo, IBM Research, USA
- Peter Eckersley, Partnership on AI, USA
- Andreas Theodorou, Umeå University, UK
- Emmanuel Arbaretier, Apsys-Airbus, France
- Yang Liu, Webank, China
- Philip Koopman, Carnegie Mellon University, USA
- Chokri Mraidha, CEA LIST, France
- Heather Roff, Johns Hopkins University, USA
- Bernhard Kaiser, ANSYS, Germany
- Brent Harrison, University of Kentucky, USA
- José M. Faria, Safe Perspective, UK
- Toshihiro Nakae, DENSO Corporation, Japan
- John Favaro, Trust-IT, Italy
- Rob Ashmore, Defence Science and Technology Laboratory, UK
- Jonas Nilsson, NVIDIA, USA
- Michael Paulitsch, Intel, Germany
- Philippa Ryan Conmy, Adelard, UK

- Ramya Ramakrishnan, Massachusetts Institute of Technology, USA
- Stefan Kugele, Technical University of Munich, Germany
- Victoria Krakovna, Google DeepMind, UK
- Richard Cheng, California Institute of Technology, USA
- Javier Ibañez-Guzman, Renault, France
- Mehrdad Saadatmand, RISE SICS, Sweden
- Alessio R. Lomuscio, Imperial College London, UK
- Rick Salay, University of Waterloo, Canada
- Jérémie Guiochet, LAAS-CNRS, France
- Sandhya Saisubramanian, University of Massachusetts Amherst, USA
- Mario Gleirscher, University of York, UK
- Guy Katz, Hebrew University of Jerusalem, Israel
- Chris Allsopp, Frazer-Nash Consultancy, UK
- Daniela Cancila, CEA LIST, France
- Vahid Behzadan, University of New Haven, USA
- Simos Gerasimou, University of York, UK
- Brian Tse, Affiliate at University of Oxford, China
- Peter Flach, University of Bristol, UK
- Gopal Sarma, Broad Institute of MIT and Harvard, USA
- Huáscar Espinoza, CEA LIST, France
- 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

As well as the additional reviewers:

- Fabio Arnez,
- Dashan Gao
- Anbu Huang

We thank Juliette Mattioli, Rodolphe Gelin and Sandhya Saisubramanian for their inspiring talks.

Finally we thank the AAAI-21 organization for providing an excellent framework for SafeAI 2021.