The AAAI-20 Workshop on Artificial Intelligence Safety (SafeAI 2020)

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[1]

We summarize the AAAI-20 Workshop on Artificial Intelligence Safety (SafeAI 2020)[2], held at the Thirty-Fourth AAAI Conference on Artificial Intelligence on February 7, New York, USA.

Introduction

Safety in Artificial Intelligence (AI) is increasingly becoming a substantial aspect 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, considering both the theoretical and practical challenges of AI safety. This view must cover a wide range of AI paradigms, considering systems that are application-specific, 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-20 Workshop on Artificial Intelligence Safety (SafeAI 2020) 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?

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

- What AI safety considerations and experiences are relevant to 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 tradeoffs with performance, affect safety?

These are the main topics of the series of SafeAI workshops. Tthey 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). This second edition was co-located with the Thirty-Fourth AAAI Conference on Artificial Intelligence on February 7, New York, USA[3].

Program

The Program Committee (PC) received 45 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, 2 talks and 15 posters, resulting in a full-paper acceptance rate of 29% and an overall acceptance rate of 67%.

The SafeAI 2019 program was organized in five thematic sessions, one keynote and two (invited) talks. We also included a short update report on the AI Safety Landscape Initiative, and poster pitches.

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: Adversarial Machine Learning

- Bio-Inspired Adversarial Attack Against Deep Neural Networks, Bowei Xi, Yujie Chen, Fei Fan, Zhan Tu and Xinyan Deng.
- Nothing to See Here: Hiding Model Biases by Fooling Post hoc Explanation Methods, Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh and Himabindu Lakkaraju.
- Adversarial Image Translation: Unrestricted Adversarial Examples in Face Recognition Systems, Kazuya Kakizaki and Kosuke Yoshida.

Session 2: Assurance Cases for AI-based Systems

- Hazard Contribution Modes of Machine Learning Components, Ewen Denney, Ganesh Pai and Colin Smith.
- Assurance Argument Patterns and Processes for Machine Learning in Safety-Related Systems, Chiara Picardi, Colin Paterson, Richard Hawkins, Radu Calinescu and Ibrahim Habli.

Session 3: Considerations for the AI Safety Landscape

- Founding the Domain of AI Forensics, Vahid Behzadan and Ibrahim Baggili.
- Exploring AI Safety in Degrees: Generality, Capability and Control, John Burden and José Hernández-Orallo.

Session 4: Fairness and Bias

- Fair Enough: Improving Fairness in Budget-Constrained Decision Making Using Confidence Thresholds, Michiel Bakker, Humberto Riveron Valdes, Duy Patrick Tu, Krishna Gummadi, Kush Varshney, Adrian Weller and Alex Pentland.
- A Study on Multimodal and Interactive Explanations for Visual Question Answering, Kamran Alipour, Jurgen P. Schulze, Yi Yao, Avi Ziskind and Giedrius Burachas.
- Models can be Learned to Conceal Unfairness from Explanation Methods, Botty Dimanov, Umang Bhatt, Mateja Jamnik and Adrian Weller.

Session 5: Uncertainty and Safe AI

- A Saddle-Point Dynamical System Approach for Robust Deep Learning, Yasaman Esfandiari, Keivan Ebrahimi, Aditya Balu, Umesh Vaidya, Nicola Elia and Soumik Sarkar.
- A High Probability Safety Guarantee with Shifted Neural Network Surrogates, Mélanie Ducoffe, Jayant Sen Gupta and Sebastien Gerchinovitz.

[3] The workshop was preceding by an independent, but related, day on the AI Safety Landscape (https://www.ai-safety.org/ai-safety-landscape). This event aimed to define an AI safety landscape: a "view" of the current needs, challenges and state of the art and the practice of this field, as a key step towards developing a body of knowledge for AI safety.

- Benchmarking Uncertainty Estimation Methods for Deep Learning With Safety-Related Metrics, Maximilian Henne, Adrian Schwaiger, Karsten Roscher and Gereon Weiss.
- PURSS: Towards Perceptual Uncertainty Aware Responsibility Sensitive Safety with ML, Rick Salay, Krzysztof Czarnecki, Maria Elli, Igancio Alvarez, Sean Sedwards and Jack Weast.

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

Keynote

- Ece Kamar (Microsoft Research AI), AI in the Open World: Discovering Blind Spots of AI.

Invited Talks

- François Terrier (Commissariat à l´Energie Atomique), Considerations for Evolutionary Qualification of Safety-Critical Systems with AI-based Components
- Sameer Singh (University of California, Irving), Evaluating and Testing Natural Language Processing Systems

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

Posters

- Simple Continual Learning Strategies for Safer Classifers, Ashish Gaurav, Sachin Vernekar, Jaeyoung Lee, Vahdat Abdelzad, Krzysztof Czarnecki and Sean Sedwards.
- "How do I fool you?": Manipulating User Trust via Misleading Black Box Explanations, Himabindu Lakkaraju and Osbert Bastani.
- Assessing the Adversarial Robustness of Monte Carlo and Distillation Methods for Deep Bayesian Neural Network Classification, Meet Vadera, Satya Narayan Shukla, Brian Jalaian and Benjamin Marlin.
- Out-of-Distribution Detection with Likelihoods Assigned by Deep Generative Models Using Multimodal Prior Distributions, Ryo Kamoi and Kei Kobayashi.
- Fair Representation for Safe Artificial Intelligence via Adversarial Learning of Unbiased Information Bottleneck, Jin-Young Kim and Sung-Bae Cho.
- SafeLife 1.0: Exploring Side Effects in Complex Environments, Carroll Wainwright and Peter Eckersley.
- (When) Is Truth-telling Favored in AI Debate?, Vojtech Kovarik and Ryan Carey.
- NewsBag: A Benchmark Multimodal Dataset for Fake News Detection, Sarthak Jindal, Raghav Sood, Richa Singh, Mayank Vatsa and Tanmoy Chakraborty.
- Algorithmic Discrimination: Formulation and Exploration in Deep Learning-based Face Biometrics, Ignacio Serna, Aythami Morales, Julian Fierrez, Manuel Cebrian, Nick Obradovich and Iyad Rahwan.
- Guiding Safe Reinforcement Learning Policies \\Using Structured Language Constraints, Bharat Prakash, Nicholas Waytowich, Ashwinkumar Ganesan, Tim Oates and Tinoosh Mohsenin.
- Practical Solutions for Machine Learning Safety in Autonomous Vehicles, Sina Mohseni, Mandar Pitale, Vasu Singh and Zhangyang Wang.
- Continuous Safe Learning Based on First Principles and Constraints for Autonomous Driving, Lifeng Liu, Yingxuan Zhu and Jian Li.
- The Incentives that Shape Behavior, Ryan Carey, Eric Langlois, Tom Everitt and Shane Legg.
- Recurrent Neural Network Properties and their Verification with Monte Carlo Techniques, Dmitry Vengertsev and Elena Sherman.
- Toward Operational Safety Verification Via Hybrid Automata Mining Using I/O Traces of AI-Enabled CPS, Imane Lamrani, Ayan Banerjee and Sandeep Gupta.

Acknowledgements

We thank all researchers who submitted papers to SafeAI 2020 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 Ece Kamar, François Terrier and Himabindu Lakkaraju for their inspiring talks.

Finally, we thank the AAAI-20 organization for providing an excellent framework for SafeAI 2020.