The IJCAI-24 Workshop on Artificial Intelligence Safety (AISafety2024)

Gabriel Pedroza[1], Xiaowei Huang[2], Xin Cynthia Chen[3], Fabio Arnez[4], Huascar Espinoza[5], José Hernández-Orallo[6], Mauricio Castillo-Effen[7], Richard Mallah[8], John McDermid[9], Andreas Theodorou[10]

[1] ANSYS, France
gabriel.pedroza@ansys.com

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

[3] ETH Zurich, Switzerland
xin.chen@inf.ethz.ch

[4] CEA LIST, France
fabio.arnez@cea.fr

[5] Chips JU, Belgium
Huascar.Espinoza@kdt-ju.europa.eu

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

[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

[10] Universitat Politècnica de Catalunya, Spain
andreas.theodorou@upc.edu

Abstract

We summarize the IJCAI-24 Workshop on Artificial Intelligence Safety (AISafety2024)[1], held at the 33nd International Joint Conference on Artificial Intelligence (IJCAI-24) on August 4, 2024 in Jeju, South Korea.

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.

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

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

The IJCAI-24 Workshop on Artificial Intelligence Safety seeks to explore new ideas in AI safety with a particular focus on addressing the following questions:

- How can we engineer trustable AI software architectures?
- Do we need to specify and use bounded morality in system engineering to make AI-based systems more ethically aligned?
- What is the status of existing approaches in ensuring AI and ML safety and what are the gaps?
- What safety engineering considerations are required to develop safe human-machine interaction in automated decision-making systems?
- What AI safety considerations and experiences are relevant from industry?
- How can we characterise or evaluate AI systems according to their potential risks and vulnerabilities?
- How can we develop solid technical visions and paradigm shift articles about AI Safety?
- How do metrics of capability and generality affect the level of risk of a system and how trade-offs can be found with performance?
- How do AI systems feature for example ethics, explainability, transparency, and accountability relate to, or contribute to, its safety?
- How to evaluate AI safety?
- How to safeguard GenAI/LLMs/ML?

These are the main topics of the series of AISafety workshops which this year have been enriched by a particular focus on Reinforcement Learning techniques, their challenges, solutions and perspectives. Overall, the series aims 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.

Program

The Program Committee (PC) received 11 submissions. Each paper was peer-reviewed by at least two PC members, by following a single-blind reviewing process. The committee decided to accept 8 full papers, resulting in an overall paper acceptance rate of 72%.

The AISafety2024 program was organized in four thematic sessions, two keynote, and one (invited) talk. 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.
- *Invited speakers* gave a 25-minute talk on a relevant topic to the workshop.
- *Keynote speakers* gave a 45-minute talk on a relevant topic to the workshop.

Presentations and papers were grouped by topic as follows:

Session 1: AI Safety Assessment: Validation Methods and Techniques

- ReLESS: A Framework for Assessing Safety in Deep Learning Systems, Anita Raja, Nan Jia, Raffi Khatchadourian
- Enhancing Autonomous Vehicle Safety through N-version Machine Learning Systems, Qiang Wen, Julio Mendonça, Fumio Machida, Marcus Völp

Session 2: AI Robustness: Adversarial Learning and Security/Privacy

- Hyper-parameter Tuning for Adversarially Robust Models, Pedro Mendes, Paolo Romano, David Garlan
- Low-Latency Privacy-Preserving Deep Learning Design via Secure MPC, Ke Lin, Yasir Glani, Ping Luo

Session 3: Safety of Generative AI: OoD and Human vs Machine Generative Detection

- Detecting Out-of-Distribution Text Using Topological Features of Transformer-Based Language Models, Anj Simmons, Andres Pollano, Anupam Chaudhuri
- The Impact of Prompts on Zero-Shot Detection of AI-Generated Text, Kouichi Sakurai, Kaito Taguchi, Yujie Gu

Session 4: AI Robustness: Resilience to Noise and Soft Errors

- Global Clipper: Enhancing Safety and Reliability of Transformer-based Object Detection Models, Qutub Syed, Michael Paulitsch, Karthik Pattabiraman, Korbinian Hagn, Fabian Oboril, Cornelius Buerkle, Kay-Ulrich Scholl, Gereon Hinz, Alois Knoll
- Neural Vicinal Risk Minimization: Noise-Robust Distillation for Noisy Labels, Hyounguk Shon, Seunghee Koh, Yunho Jeon, Junmo Kim

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

Keynotes

- Konstantin Dmitriev, The Evolvement of AI/ML Aviation Regulations and Illustration of Some Practical

Aspects through an End-to-End Certification Case Study

- Loïc Cantat, Journey and Findings of the Research Program Confiance.ai

Invited Talks

- Yonah Welker, Ability-Centered AI and Policy (Transatlantic Safety Dialogue and Designated Groups)

Acknowledgements

We thank all researchers who submitted papers to AISafety 2024 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 Nilsson, NVIDIA, USA
- Brent Harrison, University of Kentucky, USA
- Alessio R. Lomuscio, Imperial College London, UK
- 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
- Morayo Adedjouma, CEA LIST, France
- Rasmus Adler, Fraunhofer IESE, Germany
- Vahid Hashemi, Audi, Germany
- Juliette Mattioli, Thales, France
- Bonnie W. Johnson, Naval Postgraduate School, USA
- Rémi Regnier, LNE, France
- Feng Liu, Huawei, China
- 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
- Nathalie Baracaldo, IBM Research, USA
- Toshihiro Nakae, DENSO Corporation, Japan
- Gereon Weiss, Fraunhofer IKS, Germany
- Stefan Kugele, Technische Hochschule Ingolstadt, Germany
- Colin Paterson, University of York, UK
- John Burden, University of Cambridge, UK
- Fredrik Heintz, Linköping University, Sweden
- Simon Fürst, BMW Group, Germany
- Mandar Pitale, NVIDIA, USA
- Leon Kester, TNO, The Netherlands
- Bernhard Kaiser, ANSYS, Germany
- Xingyu Zhao, University of Warwick, UK
- Javier Garcia, Universidad Carlos III de Madrid, Spain

Finally, we thank the IJCAI-24 organization for providing an excellent framework for AISafety2024.