The IJCAI-23 Joint Workshop on Artificial Intelligence Safety and Safe Reinforcement Learning (AISafety-SafeRL2023)

Gabriel Pedroza[1], Xin Cynthia Chen[2], José Hernández-Orallo[3], Xiaowei Huang[4], Andreas Theodorou[5], Nikolaos Matragkas[6], Huascar Espinoza[7], Richard Mallah[8], John McDermid[9], Mauricio Castillo-Effen[10] , David Bossens[11], Bettina Koenighofer[12], Sebastian Tschiatschek[13] and Anqi Liu[14]

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

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

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

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

[5] Umeå University, Sweden
andreas.theodorou@umu.se

[6] CEA LIST, France
n.matragkas@hull.ac.uk

[7] KDT JU, Belgium
Huascar.Espinoza@kdt-ju.europa.eu

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

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

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

[11] David Bossens, University of Southampton
davidmbossens@gmail.com

[12] Bettina Koenighofer, TU Graz
bettina.koenighofer@iaik.tugraz.at

[13] Sebastian Tschiatschek, University of Vienna
sebastian.tschiatschek@univie.ac.at

[14] Anqi Liu, Johns Hopkins University
ataliu@cs.jhu.edu

Abstract

We summarize the IJCAI-23 Joint Workshop on Artificial Intelligence Safety and Safe Reinforcement Learning (AISafety-SafeRL2023)[1], held at the 32nd International Joint Conference on Artificial

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

Intelligence (IJCAI-23) on August 21-22, 2023 in Macau, China.

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.

Safe Reinforcement Learning (Safe RL) is a specialized domain within the broader field of reinforcement learning that emphasizes the importance of ensuring safety during the learning and decision-making processes. The primary objective of Safe RL is to develop algorithms and systems that can learn and make decisions without causing harm to themselves, the environment, or other entities. This encompasses avoiding physical damage, breaches of ethical standards, and violations of societal norms or legal regulations. In essence, Safe RL seeks to strike a balance between exploration and exploitation in learning, ensuring that an RL agent doesn't take actions that could lead to irreversible negative consequences, especially in critical applications like aerospace, robotics, and other safety-critical systems.

The IJCAI-23 Joint Workshop on Artificial Intelligence Safety and Safe Reinforcement Learning (AISafety-SafeRL2023) 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 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 19 submissions. Each paper was peer-reviewed by at least two PC members, by following a single-blind reviewing process. The committee decided to accept 10 full papers, resulting in an overall paper acceptance rate of 52%.

The AISafety-SafeRL2023 program was organized in five thematic sessions, two keynote and three (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.
- *Presenters* gave a 10-minute paper talk and participated in the debate slot.
- *Invited speakers* gave a 25-minute talk on a relevant topic to the workshop.
- *Contributed talk speakers* gave a 15-minute talk on a relevant topic to the workshop
- *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: Robustness of AI via OoD and Unknown-Unknowns Dectection

- Diffusion Denoised Smoothing for Certified and Adversarial Robust Out Of Distribution, Nicola Franco, Daniel Korth, Jeanette Miriam Lorenz, Karsten Roscher and Stephan Günnemann

- Unsupervised Unknown Unknown Detection in Active Learning, Prajit T. Rajendran, Huascar Espinoza, Agnes Delaborde and Chokri Mraidha

Session 2: AI Robustness, Adversarial Attacks and Reinforcemnt Learning

- PerCBA: Persistent Clean-label Backdoor Attacks on Semi-Supervised Graph Node Classification, Xiao Yang, Gaolei Li, Chaofeng Zhang, Meng Han and Wu Yang
- Distribution-restrained Softmax Loss for the Model Robustness, Chen Li, Hao Wang, Jinzhe Jiang, Xin Zhang, Yaqian Zhao and Weifeng Gong
- Fear Field: Adaptive constraints for safe environment transitions in Shielded Reinforcement Learning, Haritz Odriozola-Olalde, Nestor Arana, Arexolaleiba, Maider Zamalloa, Jon Perez, Cerrolaza, Jokin Arozamena and Rodríguez

Session 3: AI Governance and Policy/Value Alignment

- An open source perspective on AI and alignment with the EU AI Act Diego Calanzone, Andrea Coppari, Riccardo Tedoldi, Giulia Olivato and Carlo Casonato

Session 4: SafeRL

- Yanan Sui: Embodied safe optimization for the restoration of human motor functions
- Thiago Simao: Ensuring the offline reliability and online safety of reinforcement learning agents
- Filip Cano: Search-based Testing of Reinforcement Learning
- Martin Kurezca: Monte Carlo Tree Search with Function Approximation for Risk-constrained Planning and Reinforcement Learning
- Ruoqi Zhang: Risk-sensitive Actor-free Policy via Convex Optimisation
- Weiye Zhao: State-wise Constrained Policy Optimization

Session 5: AI Trustworthiness, Explainability and Testing

- Empirical Optimal Risk to Quantify Model Trustworthiness for Failure Detection, Shuang Ao, Stefan Rueger and Advaith Siddharthan
- Weight-based Semantic Testing Approach for Deep Neural Networks, Amany Alshareef, Nicolas Berthier, Sven Schewe and Xiaowei Huang
- AI for Safety: How to use Explainable Machine Learning Approaches for Safety Analyses Iwo Kurzidem, Simon Burton and Philipp

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

Keynotes

- Paul Lukowicz, Safety risks of AI: Intelligence, Complexity and Stupidity
- François Terrier, No Trust without regulation! European challenge on regulation, liability and standards for trusted AI

Invited Talks

- Yanan Sui: Embodied safe optimization for the restoration of human motor functions
- Thiago Simao: Ensuring the offline reliability and online safety of reinforcement learning agents

Acknowledgements

We thank all researchers who submitted papers to AISafety 2023 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
- 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
- Morayo Adedjouma, CEA LIST, 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 IKS, 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
- Bernhard Kaiser, ANSYS, Germany

Finally we thank the IJCAI-23 organization for providing an excellent framework for AISafety-SafeRL2023.