SafeAI 2022

The AAAI-22 Workshop on Artificial Intelligence Safety

Feb 28th - Mar 1st, 2022
Virtual

Gabriel Pedroza, CEA LIST, France
José Hernández-Orallo, Universitat Politècnica de València, Spain
Xin Cynthia Chen, University of Hong Kong, China
Xiaowei Huang, University of Liverpool, UK
Huáscar Espinoza, ECSEL JU, Belgium
Mauricio Castillo-Effen, Lockheed Martin, USA
Seán Ó hÉigeartaigh, University of Cambridge, UK
Richard Mallah, Future of Life Institute, USA
John McDermid, University of York, UK

Opening Remarks

*The main interest of SafeAI 2022 is to explore new ideas on **AI safety** by looking holistically at theoretical and practical, short-term and long-term, perspectives, jointly with the ethical and legal issues, to build trustable intelligent autonomous machines.*

- As SafeAI aims at bringing together multiple perspectives, it's probably possible to harshly criticise any paper here today… Most likely anyone has missed some important issue…
- So, please do be critical, but temper your criticism with constructive discussions!

Program (See Website for Details)

SafeAI

Day 1: Scheduled on **Feb 28, 2022** from **13:00 to 18:10 UTC** [**5:00-10:10 PST**]

Location: AAAI Virtual Venue, Red Building Room 5.

Time (UTC)	Description
13:00-13:05	Welcome and Introduction – Chair: Gabriel Pedroza (CEA List)
13:05-13:50	Keynote 1: Matthew Dwyer (University of Virginia), *Distribution-aware Test Adequacy for Neural Networks*
13:50-14:00	Short Break
14:00-15:30	**Special Session 1: *EnnCore* – Chair: Lucas Cordeiro (University of Manchester)** **EnnCore** addresses the fundamental problem of guaranteeing safety, transparency, and robustness in neural-based architectures. DETAILED ENNCORE PROGRAM...
15:30-16:00	Coffee Break
16:00-16:25	Invited Talk 1: Shiri Dori-Hacohen (University of Connecticut), *Quantifying Misalignment Between Agents*

Day 2: Scheduled on **Mar 1, 2022** from **8:00 to 17:15 UTC** [**0:00-9:15 PST**]

Location: AAAI Virtual Venue, Red Building Room 5.

Time (UTC)	Description
8:00-8:25	Invited Talk 2: Roel Dobbe (TU Delft), *A System Safety Perspective for Developing and Governing Artificial Intelligence*
8:25-9:35	**Technical Session 3: *Robustness and Uncertainty* – Chair: Xin Cynthia Chen (University of Hong Kong)** – Efficient Adversarial Sequence Generation for RNN with Symbolic Weighted Finite Automata, Mingjun Ma, Dehui Du, Yuanhao Liu, Yanyun Wang and Yiyang Li. – A Study on Mitigating Hard Boundaries of Decision-Tree-based Uncertainty Estimates for AI Models, Pascal Gerber, Lisa Jöckel and Michael Kläs. – Quantifying the Importance of Latent Features in Neural Networks, Amany Alshareef, Nicolas Berthier, Sven Schewe and Xiaowei Huang. – Maximum Likelihood Uncertainty Estimation: Robustness to Outliers, Deebul Nair, Nico Hochgeschwender and Miguel Olivares-Mendez. – Debate Panel – Paper Discussants: Xiaowei Huang (University of Liverpool), Mauricio Castillo-Effen (Lockheed Martin)

Some Additional Information

- SafeAI 2022 Best Paper Award.
- Proceedings is freely available at CEUR-WS: http://ceur-ws.org/Vol-3087/
  (URL is available at the SafeAI website)
- Presentations will be available on the website very soon
- We hope you enjoy SafeAI 2022!