SafeAI 2021

The AAAI-21 Workshop on Artificial Intelligence Safety

Feb 8th, 2021
Virtual

Huáscar Espinoza, Commissariat a'l'Energie Atomique, 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, UK

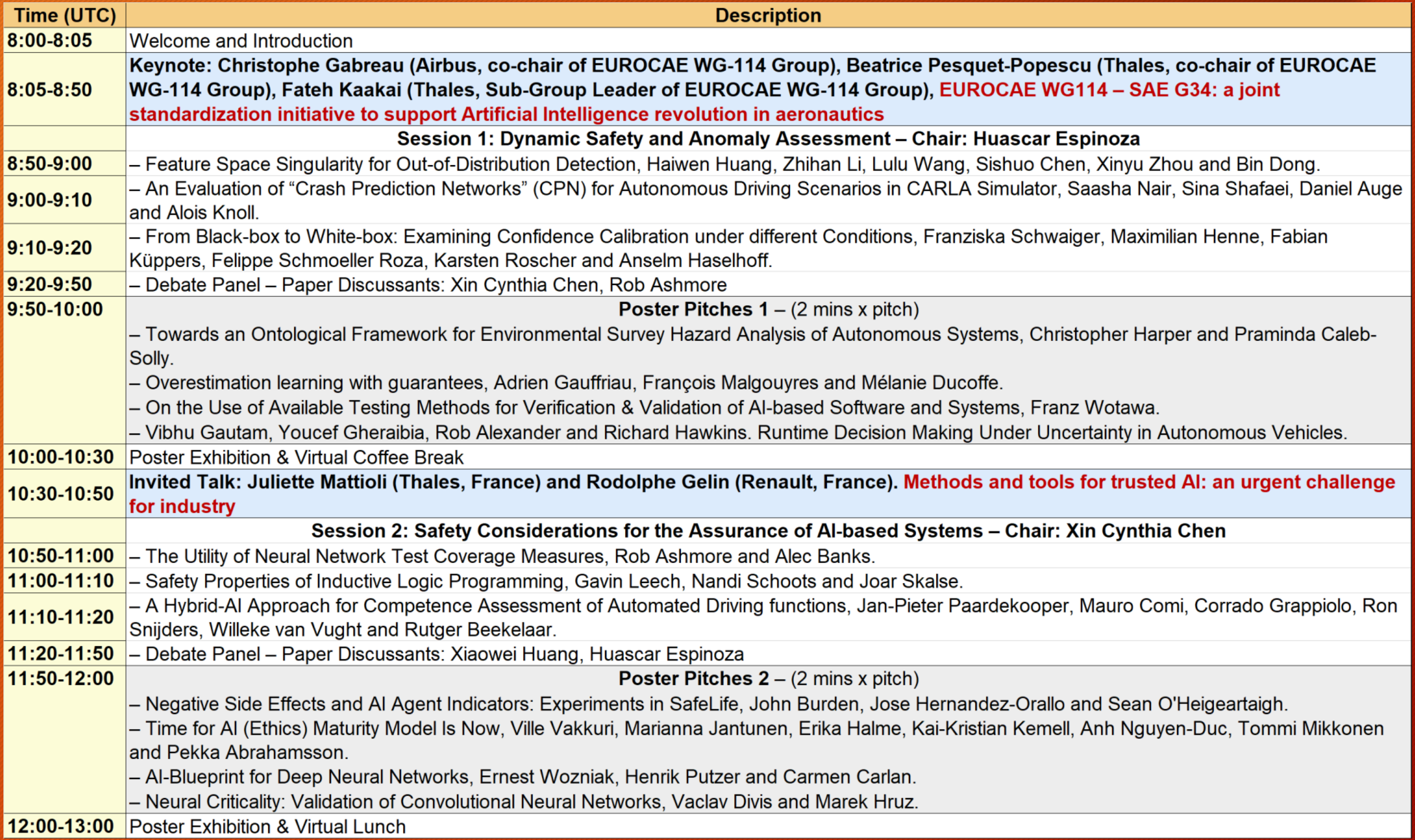
Opening Remarks

SafeAI

*The main interest of SafeAI 2021 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.*

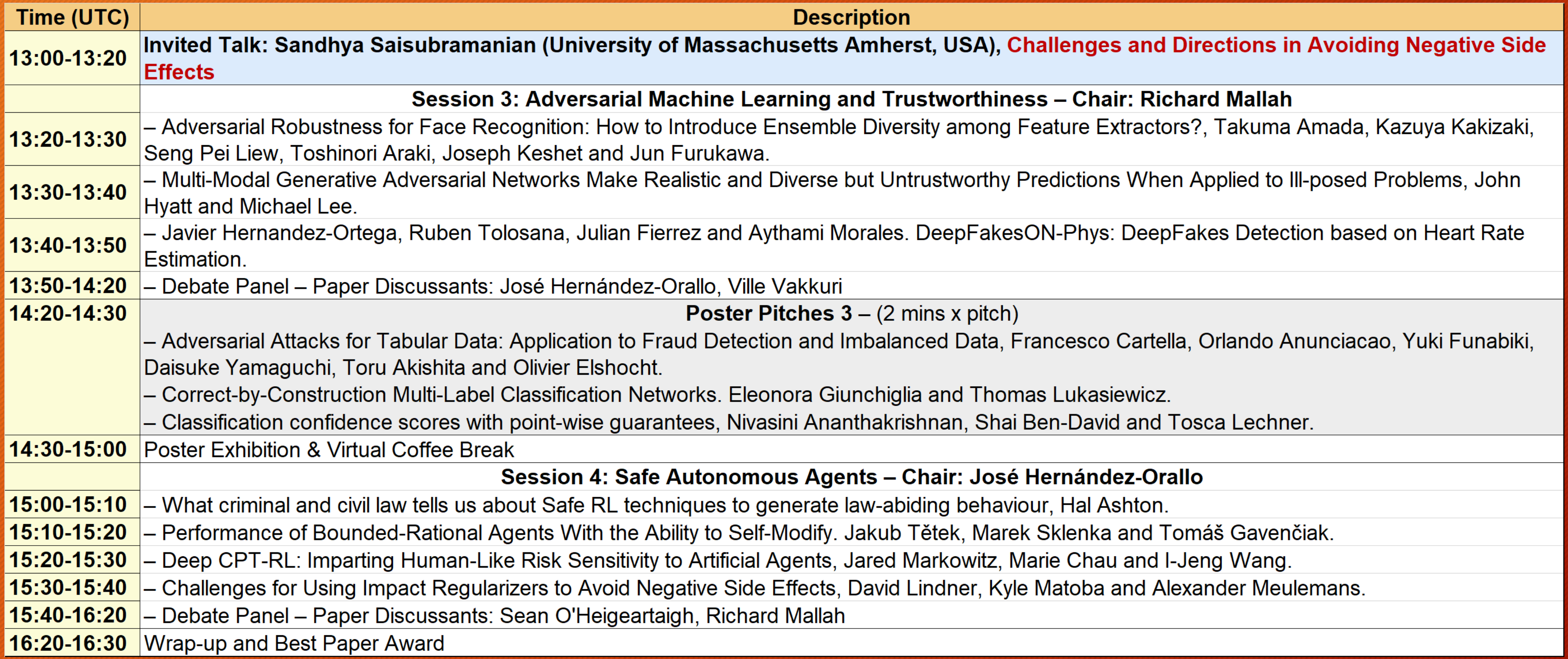
- As SafeAI aims at bringing together multiple perspectives, it’s probably possible to harshly criticise any paper here today… Most likely anyone has missed some important issue…
- So, please do be critical, but temper your criticism with constructive discussions!

Program (Morning)



Time (UTC)	Description
8:00-8:05	Welcome and Introduction
8:05-8:50	**Keynote: Christophe Gabreau (Airbus, co-chair of EUROCAE WG-114 Group), Beatrice Pesquet-Popescu (Thales, co-chair of EUROCAE WG-114 Group), Fateh Kaakai (Thales, Sub-Group Leader of EUROCAE WG-114 Group), EUROCAE WG114 – SAE G34: a joint standardization initiative to support Artificial Intelligence revolution in aeronautics**
	Session 1: Dynamic Safety and Anomaly Assessment – Chair: Huascar Espinoza
8:50-9:00	– Feature Space Singularity for Out-of-Distribution Detection, Haiwen Huang, Zhihan Li, Lulu Wang, Sishuo Chen, Xinyu Zhou and Bin Dong.
9:00-9:10	– An Evaluation of “Crash Prediction Networks” (CPN) for Autonomous Driving Scenarios in CARLA Simulator, Saasha Nair, Sina Shafaei, Daniel Auge and Alois Knoll.
9:10-9:20	– 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.
9:20-9:50	– Debate Panel – Paper Discussants: Xin Cynthia Chen, Rob Ashmore
9:50-10:00	**Poster Pitches 1** – (2 mins x pitch) – 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. – Vibhu Gautam, Youcef Gheraibia, Rob Alexander and Richard Hawkins. Runtime Decision Making Under Uncertainty in Autonomous Vehicles.
10:00-10:30	Poster Exhibition & Virtual Coffee Break
10:30-10:50	**Invited Talk: Juliette Mattioli (Thales, France) and Rodolphe Gelin (Renault, France). Methods and tools for trusted AI: an urgent challenge for industry**
	Session 2: Safety Considerations for the Assurance of AI-based Systems – Chair: Xin Cynthia Chen
10:50-11:00	– The Utility of Neural Network Test Coverage Measures, Rob Ashmore and Alec Banks.
11:00-11:10	– Safety Properties of Inductive Logic Programming, Gavin Leech, Nandi Schoots and Joar Skalse.
11:10-11:20	– 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.
11:20-11:50	– Debate Panel – Paper Discussants: Xiaowei Huang, Huascar Espinoza
11:50-12:00	**Poster Pitches 2** – (2 mins x pitch) – 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.
12:00-13:00	Poster Exhibition & Virtual Lunch

Program (Afternoon)



Time (UTC)	Description
13:00-13:20	**Invited Talk: Sandhya Saisubramanian (University of Massachusetts Amherst, USA), Challenges and Directions in Avoiding Negative Side Effects**
	Session 3: Adversarial Machine Learning and Trustworthiness – Chair: Richard Mallah
13:20-13:30	– 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.
13:30-13:40	– Multi-Modal Generative Adversarial Networks Make Realistic and Diverse but Untrustworthy Predictions When Applied to Ill-posed Problems, John Hyatt and Michael Lee.
13:40-13:50	– Javier Hernandez-Ortega, Ruben Tolosana, Julian Fierrez and Aythami Morales. DeepFakesON-Phys: DeepFakes Detection based on Heart Rate Estimation.
13:50-14:20	– Debate Panel – Paper Discussants: José Hernández-Orallo, Ville Vakkuri
14:20-14:30	**Poster Pitches 3** – (2 mins x pitch) – 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.
14:30-15:00	Poster Exhibition & Virtual Coffee Break
	Session 4: Safe Autonomous Agents – Chair: José Hernández-Orallo
15:00-15:10	– What criminal and civil law tells us about Safe RL techniques to generate law-abiding behaviour, Hal Ashton.
15:10-15:20	– Performance of Bounded-Rational Agents With the Ability to Self-Modify. Jakub Tětek, Marek Sklenka and Tomáš Gavenčiak.
15:20-15:30	– Deep CPT-RL: Imparting Human-Like Risk Sensitivity to Artificial Agents, Jared Markowitz, Marie Chau and I-Jeng Wang.
15:30-15:40	– Challenges for Using Impact Regularizers to Avoid Negative Side Effects, David Lindner, Kyle Matoba and Alexander Meulemans.
15:40-16:20	– Debate Panel – Paper Discussants: Sean O'Heigeartaigh, Richard Mallah
16:20-16:30	Wrap-up and Best Paper Award

Some Additional Information

SafeAI

- Voting for SafeAI 2021 Best Paper Award:

  www.menti.com – Code: **57 59 18 6**

- Proceedings is freely available at CEUR-WS: http://ceur-ws.org/Vol-2808/

  (URL is available at the SafeAI website)

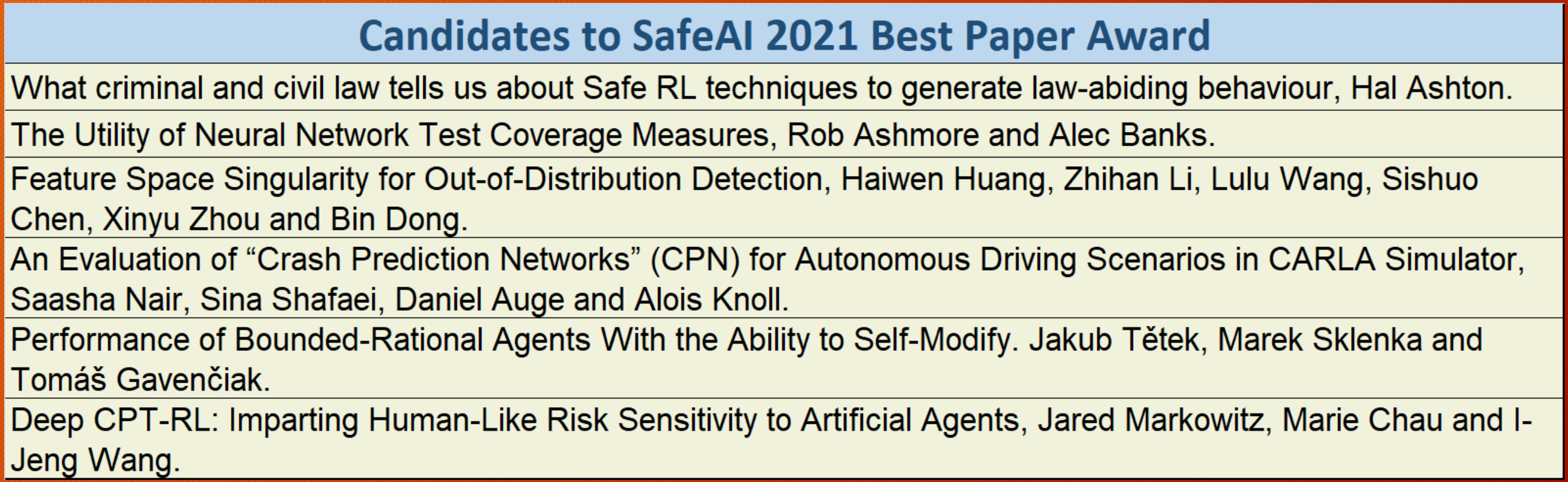
- Presentations will be available on the website very soon

- We hope you enjoy SafeAI 2021!

Wrap Up

SafeAI

- Voting for SafeAI 2021 Best Paper Award:
  www.menti.com – Code: **57 59 18 6**
- Proceedings is freely available at CEUR-WS: http://ceur-ws.org/Vol-2808/
  (URL is available at the SafeAI website)



Candidates to SafeAI 2021 Best Paper Award
What criminal and civil law tells us about Safe RL techniques to generate law-abiding behaviour, Hal Ashton.
The Utility of Neural Network Test Coverage Measures, Rob Ashmore and Alec Banks.
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.
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.

- Join the CLAIS (Consortium of the Landscape on AI Safety): www.clais.org