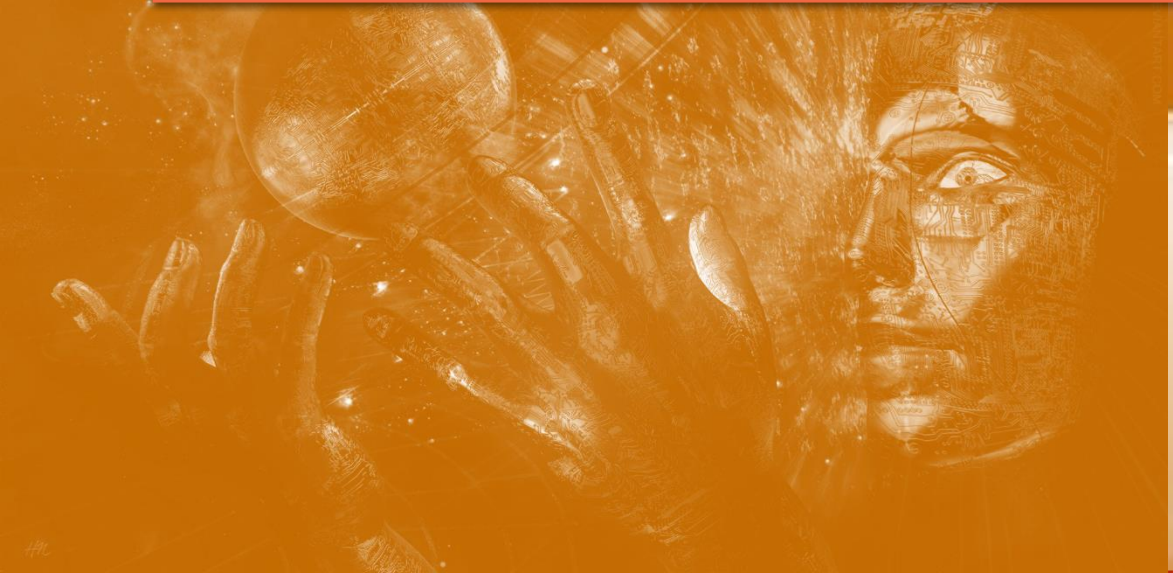
SafeAI 2020

The AAAI-20 Workshop on Artificial Intelligence Safety

Feb 7th, 2020
NYC, USA

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

The main interest of SafeAI 2020 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.*

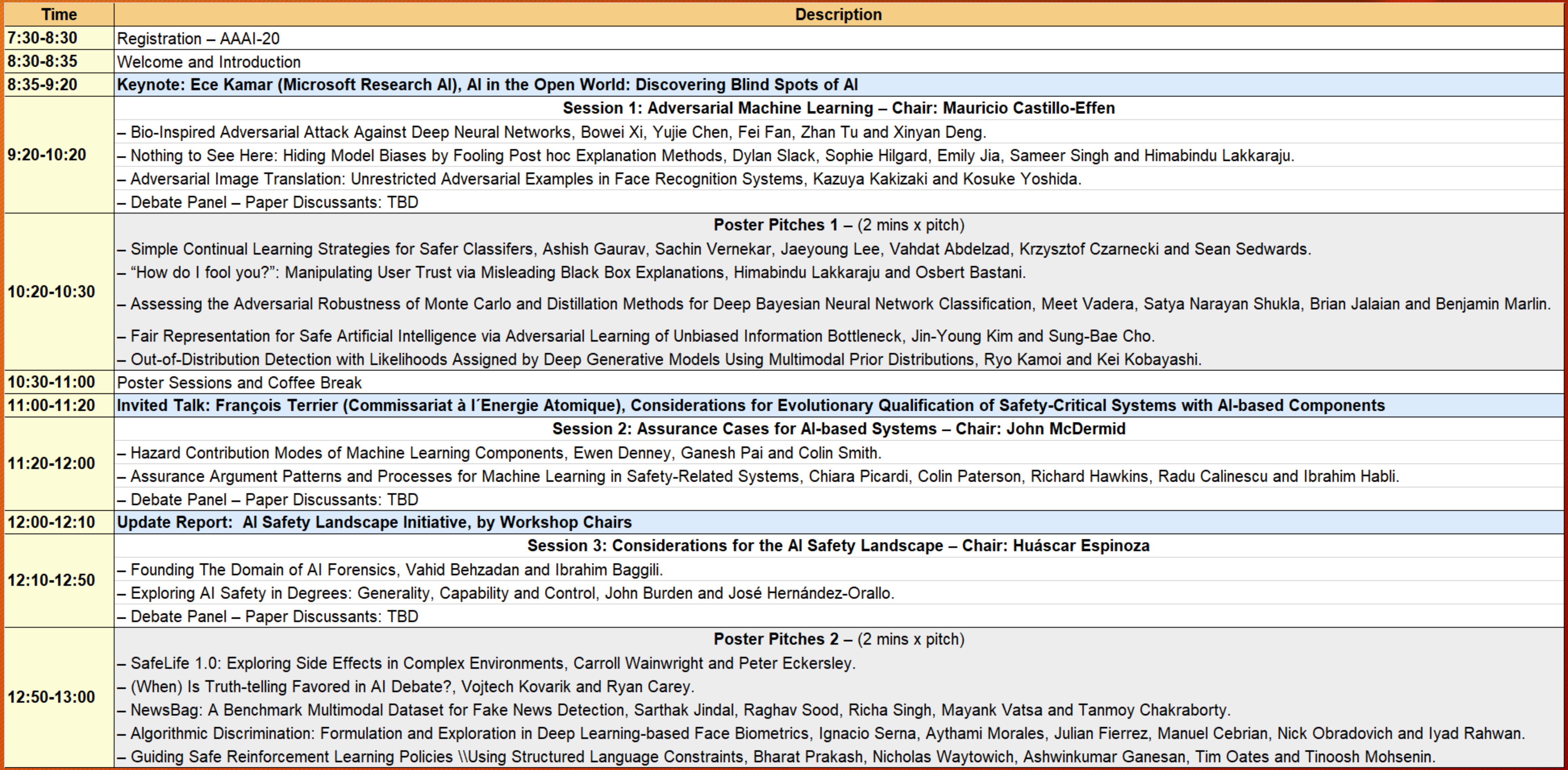


Opening Remarks

- 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!
- The AI Safety community must be voraciously interdisciplinary if it is to be useful.

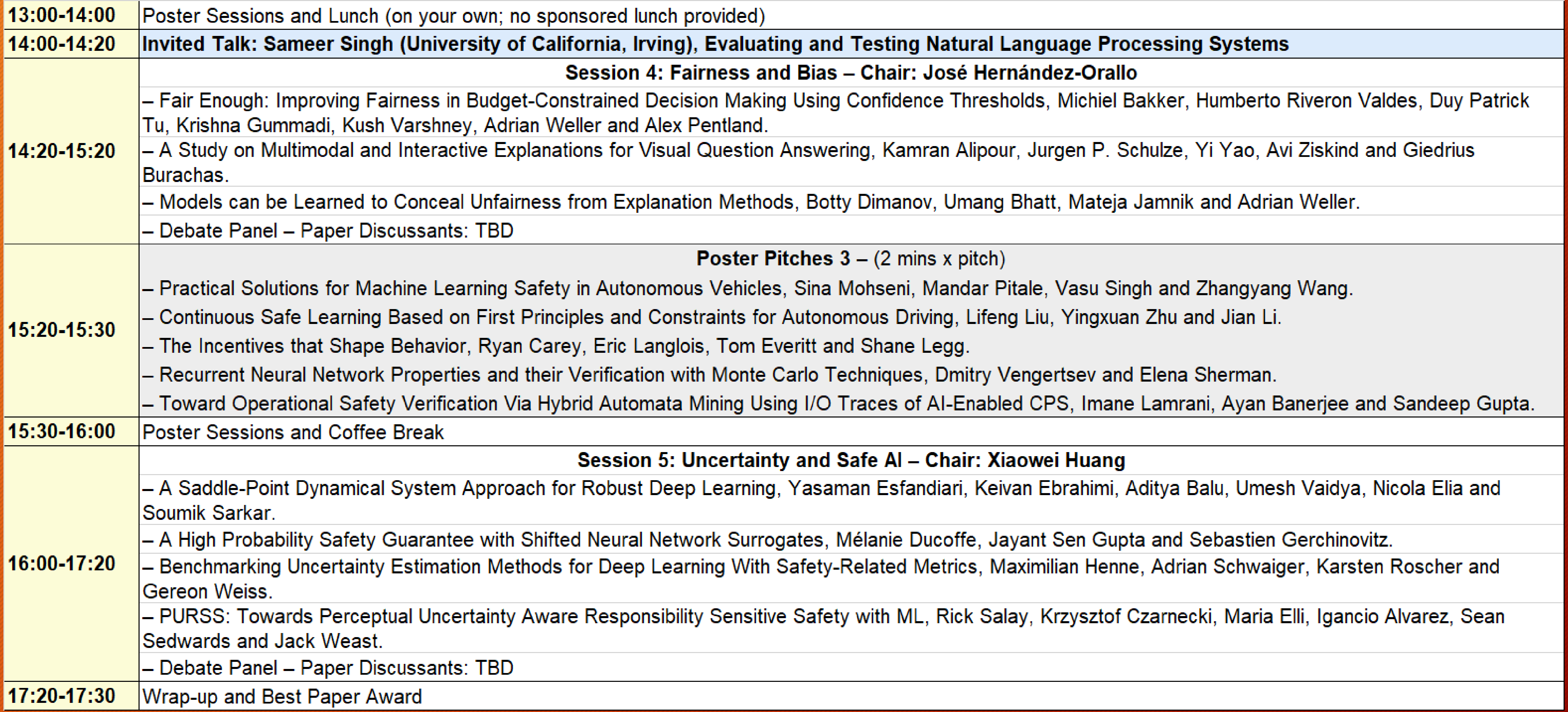
Program (Morning)

SafeAI



Time	Description
7:30-8:30	Registration – AAAI-20
8:30-8:35	Welcome and Introduction
8:35-9:20	**Keynote: Ece Kamar (Microsoft Research AI), AI in the Open World: Discovering Blind Spots of AI**
9:20-10:20	**Session 1: Adversarial Machine Learning – Chair: Mauricio Castillo-Effen**
	– 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.
	– Debate Panel – Paper Discussants: TBD
10:20-10:30	**Poster Pitches 1** – (2 mins x pitch)
	– 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.
	– Fair Representation for Safe Artificial Intelligence via Adversarial Learning of Unbiased Information Bottleneck, Jin-Young Kim and Sung-Bae Cho.
	– Out-of-Distribution Detection with Likelihoods Assigned by Deep Generative Models Using Multimodal Prior Distributions, Ryo Kamoi and Kei Kobayashi.
10:30-11:00	Poster Sessions and Coffee Break
11:00-11:20	**Invited Talk: François Terrier (Commissariat à l´Energie Atomique), Considerations for Evolutionary Qualification of Safety-Critical Systems with AI-based Components**
11:20-12:00	**Session 2: Assurance Cases for AI-based Systems – Chair: John McDermid**
	– 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.
	– Debate Panel – Paper Discussants: TBD
12:00-12:10	**Update Report: AI Safety Landscape Initiative, by Workshop Chairs**
12:10-12:50	**Session 3: Considerations for the AI Safety Landscape – Chair: Huáscar Espinoza**
	– 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.
	– Debate Panel – Paper Discussants: TBD
12:50-13:00	**Poster Pitches 2** – (2 mins x pitch)
	– 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.

Program (Afternoon)



13:00-14:00	Poster Sessions and Lunch (on your own; no sponsored lunch provided)
14:00-14:20	**Invited Talk: Sameer Singh (University of California, Irving), Evaluating and Testing Natural Language Processing Systems**
14:20-15:20	**Session 4: Fairness and Bias – Chair: José Hernández-Orallo**
	– 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.
	– Debate Panel – Paper Discussants: TBD
15:20-15:30	**Poster Pitches 3 –** (2 mins x pitch)
	– 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.
15:30-16:00	Poster Sessions and Coffee Break
16:00-17:20	**Session 5: Uncertainty and Safe AI – Chair: Xiaowei Huang**
	– 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.
	– 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.
	– Debate Panel – Paper Discussants: TBD
17:20-17:30	Wrap-up and Best Paper Award

Some Additional Information

SafeAI

- Voting for SafeAI 2020 Best Paper Award:

  www.menti.com – Code: **37 05 08**

- Proceedings will be freely available at CEUR-WS:

  URL will be soon available at the SafeAI website

- Presentations will be available on the website very soon

- We hope you enjoy SafeAI 2020!