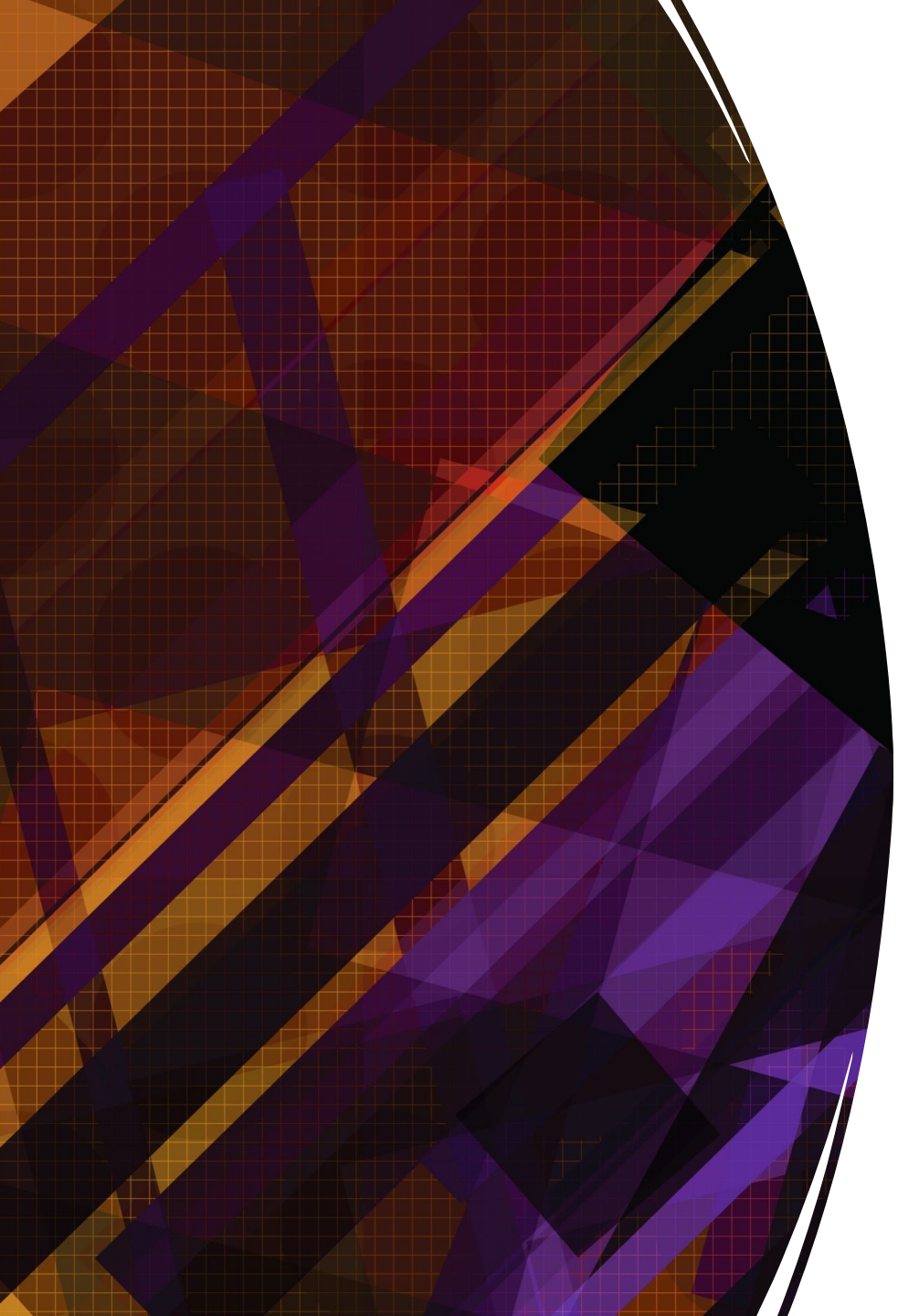

SAFETY AND SECURITY OF ML-ENABLED CPS

Session Chair: Domenico Cotroneo

Scribe: Carl Landwehr

WE HAD TWO PRESENTATIONS

- Karthik Pattabiraman, UBC
  - Building Error-resilient and Attack-resilient ML-enabled CPS
- Mahsa Ghasemi, Purdue
  - No-Regret Learning for Trustworthy Online Decision-Making

KARTHIK

- General objective: *Resiliency for both faults and attacks* (w/o human intervention)
- Fault/Attack models
  - Soft errors
  - Training data faults
  - Adversarial patch attacks
- Specific goal
  - Improve the Silent Data Corruption Coverage (SDC)
  - Reduce the overhead

KARTHIK

- Key (interesting) Idea
  - Transform Critical Faults into Benign Faults, via Selective Range Restriction in Hidden Layers
- Also: use of diverse ensembles of ML systems to mitigate faults
  - Last August, Al Avizienis proposed to Brian Randell and me via email a scheme for dealing with «hallucinations» based on comparing the outputs of independently developed ML systems. The systems would all generate hallucinations, but perhaps they would differ.
  - To me, achieving independence would require independent sets of training data, which would be a challenge, so I did not endorse the appaorch. Brian agreed. Maybe I was too hasty!

MAHSA GHASEMI, PURDUE

- The work studies No-Regret Learning for Trustworthy Online Decision-Making in Stackelberg Games
- Stackelberg game is a leader-follower game - e.g., when a leader firm decides to adjust its prices, and then a follower firm decides whether to do the same or something else
- Regret is what game player experiences when, after making a decision, the player sees the other player's action and realizes that they could have had a bigger reward if they had made a different choice (i.e. the choice they made was not, in retrospect, optimal.
- Online decision making means the game is played dynamically, participants learn over time
- Trustworthy means .... ?

MAHSA

- The idea (I think) is to find algorithms so that the leader can gradually discover/model the follower's utility function and in the process accumulate minimal regret
- One of the papers cited in the presentation presents an application of the method to park rangers charged with protecting a variety of animals in a refuge. The rangers must allocate resources to different regions of the park, knowing the distribution of animals but not the utlity function of the poachers (i.e. which animals the poachers are after)
- No doubt there is an application of this model to CPS systems and their attackers, but I haven't figured it out

DISCUSSION PERIOD

- Following the paper presentations, a lively discussion of what exactly we mean by "Trustworthy" and other dependability-related terms ensued
- At the conclusion of the session, it seemed there was some appetite among the group to consider expanding the set of dependability attributes from those identified in earlier WG 10.4 discussions and to pursue quantification of the new attributes