What is missing in mission-critical systems

Session Chair: Elias Duarte

Rapporteur: António Casimiro

Crafting ML Components in Safety-Critical Systems

Andrea Bondavalli

- Landscape:
  - Complexity (sophistication) of functions + unpredictable environments + not well understood technologies
  - Standards: ISO 26262 + SOTIF + ISO PAS 8800
  - Complex landscape of standards and regulation concerning AI
- Work on ML classifiers:
  - Insight: integrate fail-controlled mechanisms, rather than striving for perfect accuracy
  - Kind of fail-aware classifier (uncertain predictions are transformed into omissions)
  - Set of metrics to evaluate these fail-controlled classifiers

Crafting ML Components in Safety-Critical Systems
Andrea Bondavalli

- Multiple possible architectures using these classifiers
  - Wrapper, stacking, voting, recovery block, etc
- Experiments with binary and multiclass classifiers, with a tabular and an image dataset
  - Results highlight the positive impact of adopting the different architectures, in comparison to using the individual ML component alone
  - They also highlight the diversity among the different strategies, as each of them leads to different results

- **Key take-away**: Fail-Controlled Classifiers provide a safer alternative for classifiers in safety critical systems

What's Missing from Computer-based System Safety?

Phil Koopman

- Provocation: Safety standards are broken
- What is “Safety”? Different meanings for different persons
- When using ML, the environment is no longer defined
  - Some accident examples as illustrations

- Risk definition broken?
  - Counting deaths numbers is not sufficient – depends on who dies
  - Examples that the concept has several facets

What's Missing from Computer-based System Safety?

Phil Koopman

- New definitions for:
  - Loss
  - Risk
  - Safety constraint
  - Safety engineering
  - Safety case
  - Acceptable safety

Key take-away: Some existing definitions do not capture important aspects of the problem (they ignore that the human driver is not there), so new definitions are provided

Missed something? https://philkoopman.substack.com/p/keynote-talk-understanding-self-driving

Safety-Critical Systems: Human Factors Are Back in Full Force, Henrique Madeira

- Example: Breast cancer diagnosis
  - Good systems create blind trust, which may be problematic
  - Perception of “good” behaviour (accuracy between 90% and 99.9%) is very far from the really needed accuracy (several nines)
- Example: Software development
  - AI tools that help developing code, reducing coding time and efforts
  - But the reality is different, as the AI tools (LLMs) introduce too many bugs and security vulnerabilities. Understanding the generated code is difficult.
  - Human factors will define the future of software development
- Example: Neuro Software Engineering
  - Using neurophysiological methods for software engineering
  - iReview Experiment: using sensors to measure human reactions and the degree to which someone is understanding the code
- **Key take-away**: Human factors will become increasingly important

Discussion

- HM - Code reviewing is used nowadays, before testing, which justifies the relevance of the iReview concept
- PK - AI-assisted driving is not at all safer. It is a commodity feature, not a safety feature
- PK - How to describe/evaluate the coverage instead of simply counting the miles travelled? There is on going work on this.
- AB - What about the resilience of models to backdoors. Tests were done with out of distribution data, not with poisoned data