*Pașcan Adrian*

*248-2*

*18/11/2021*

**Paper Report on**

*Safe Artificial Intelligence and Formal Methods*

Emil Vassev

*International Symposium on Leveraging Applications of Formal Methods*

Springer, Cham, 2016 <https://ulir.ul.ie/bitstream/handle/10344/5407/Vassev_2016_safe.pdf?sequence=2>

1. Subject

* Use of formal methods for AI safety, both in life and technology.

1. Motivation

* In the near future, AI will go beyond the original human intensions, implying a level of artificial awareness. This, in turn, requires knowledge representation as a formal specification.

1. Results

* AI requirements are derived in a 4-stage process: hazard identification, hazard analysis, identifying safety capabilities and requirements derivation.
* AI is about to reach technological singularity by being able to decide which hazards are not hazards anymore.
* Formal methods: do not eliminate testing, but complement it; help automatic test case generation through formal specifications; add on safety, do not guarantee it.
* A good use of safety measures imposed by formal methods may lead to a system with up to 99% safety guarantee. These measures are: early detection of safety flaws; high quality requirements improve their design & implementation; help derivation and generation of efficient test cases.
* Human be haviour and principles cannot be formalized.
* Verification tools for fuzzy control systems are not efficient due to the huge state-explosion problem.
* Deductive guarantees prove simple safety properties and are prone to resolution refutation.
* Probabilistic guarantees check more complex properties with some degree of uncertainty and may bring some extra confidence to safety properties that cannot be handled otherwise.
* Abstraction is the most efficient solution to the state-explosion problem. State space is reduced by aggregating state transitions into coarser-grained state transitions, but the granularity of the system may be get to a point where it no longer adequately represents that system.
* Global safety invariants are goal, behaviour, interaction and resource invariants.
* Contemporary model checking is suitable for finite, yet small spaces (or larger spaces with the use of symbolic/probabilistic model checking).
* New verification toolset should use stabilization science (and stability analysis), code generation & analysis techniques for efficient test-case generation, automatic test case generation and high-performance computing parallelization.

1. Conclusion

* Uncertainty is part of the development process and may be covered in a wide range by deductive and probabilistic guarantees.
* The Contemporary formal verification toolset is not powerful enough to guarantee safety. New and enhanced toolset includes better automated reasoning and model checking, plus new verification techniques based on stabilization science, test-case generation and simulation, parallelized using high-performance computing.