TOWARDS SAFE-BY-DESIGN ARTIFICIAL INTELLIGENCE SYSTEMS
Safe-by-Design is the concept of applying methods to minimize hazards early in the design process.
AI-based systems help to solve challenging issues from cancer detection to image understanding and natural language processing.

On 8 April 2019, the High-Level Expert Group on AI (HLEG AI) presented Ethics Guidelines* for Trustworthy Artificial Intelligence: “AI systems need to be resilient and secure. They need to be safe…”

CHALLENGES

- **How to develop AI-based systems?**
  - Lack of standards describing development life-cycle of AI-based systems

- **How to ensure safety of AI-based systems?**
  - Lack of methodological support for safety analysis of AI-based systems
Propose a methodology for coupling the development of AI-based systems with safety analysis.
METHODOLOGY

PHASE 1

1b. Safety Management

PHASE 2

1a. Objectives & requirements

2. Functional decomposition & allocation

2.3a. Safety data gathering & exploration

HARA (ISO26262)
PHA, FHA, SHA
HazOp
STPA

2.3b. Safety analysis & risk assessment

PHA,
FHA,
SHA
HazOp
STPA

PHASE 3

3a. AI architecture evaluation & assessment towards safety requirements

Model, target variables, obj. functions

Refined safety models & objective functions satisfying safety req.

3b. AI architecture evaluation & assessment towards safety requirements

Training Data

PHASE 4

4. Issues diagnosis

PHASE 5

5. System Integration, Test & Evolution

end

Is model good enough?

NO

YES

Is AI-based component?

NO

YES

2.2c. Conventional design, implementation, test, verification, validation, evolution

2.2b. Data gathering & exploration

2.2a. AI component building

Problem formulation

AI technique selection/dev.

Design of objective func.

AI model development

3a. AI architecture interpretation, evaluation & assessment

Refined safety models & objective functions satisfying safety req.

Safety goals, objectives, requirements
OPEN QUESTIONS

• **Accuracy** and **completeness** of safety analysis:
  • It is performed on data that may not be representative of the infinite number of scenarios the systems can face.

• **Controllability** of AI systems is limited as the human is no longer in the loop.
FURTHER WORK

• Refine the methodology to more particular AI techniques (e.g. deep learning, cognitive computing).

• Adapt the classical safety methods (PHA, HARA, SHA, HazOp, etc.) to the context of specific AI techniques.

• Validate the methodology through application on industry-relevant case studies.