The Institution of Engineering and Technology

Artificial intelligence and functional safety

This paper provides a top-level summary to support decision making on the use of artificial intelligence (AI) in safety related systems. We aim to highlight the behaviour and risks associated with AI and to consider various techniques and measures used during the engineering lifecycle.

We define AI as software used to solve problems that it was not specifically programmed for. Current technologies have only achieved relatively low levels of narrow AI. Functional Safety[1] is the part of Equipment Under Control safety that relates to the correct functioning of electrical/electronic/programmable electronic safety-related systems.

AI is an enabling technology for autonomous systems. Its use in safety-critical product development is increasing significantly and delivering benefits for users. We have focused on 10 key pillars.

IET AI-Safety policy position

1. Data
2. Legal and ethical considerations
3. Learning
4. Verification and validation
5. Security
6. Algorithmic behaviours
7. Human factors
8. Dynamic hazards and safety arguments
9. Maintenance and operation
10. Specification

Foundation of AI-safety:
- **What is AI-safety?**
- **How can AI impact safety culture?**
- **AI systems in mixed integrity environments.**

[1] BSI, Functional safety of electrical/electronic/programmable electronic safety related systems. Part 4. BS EN 61508-4:2010.

The Institution of Engineering and Technology

1 Data

The categorisation and use of data that incorporate AI behaviour is loosely divided into input, training, test and experience. To ensure the safety system achieves the required performance, data must be sufficiently: independent; reliable (have equal integrity to the safety function); diverse; and comprehensive.

2 Legal and ethical considerations

AI and its associated data bring challenges that may not arise in traditional systems. Guidance should focus on a broad range of societal areas of concern. These include (but are not limited to): privacy; accountability; transparency and explainability; fairness and non-discrimination; human control of technology; and the promotion of human values.

3 Learning

Machine learning (ML) uses an adaptive model trained on data to produce its own model (algorithm) of the problem domain by extracting relationships and knowledge from data. The main categories of AI learning are supervised, unsupervised and reinforcement learning. The choice of learning technique depends on the problem to be solved and the available data.

4 Verification and validation (V&V)

AI software is too complex for detailed requirements, against which to verify the behaviour. ML, the main behaviour source, is exposed to a learning environment, which cannot be accurately defined. Emergent behaviour may be desirable to support system adaptations if confidence in safety is maintained. Current V&V techniques do not yet provide the same assurance as traditional techniques. Some approaches set defined safety boundary conditions for the AI to operate within.

5 Security

A system has to be secure to be safe and hence it must be considered throughout its lifecycle. This includes design, training, deployment, operation, maintenance and retirement.

6 Algorithmic behaviours

Unlike traditional software, there is no defined model available for interrogation with AI systems, so theoretical behaviours cannot be verified in the same way. Key questions can help determine whether AI can deliver the required integrity level. These include, but are not limited to: Has an appropriate algorithm type been selected? Is it possible to explain its output? Does it support failure identification and demonstrate resilience?

7 Human factors (HF)

The implementation of AI in safety critical applications is likely to require the re-evaluation of tried and tested HF management philosophies. For the design phase, lifecycle challenges include that AI systems cannot interpret poorly specified attributes, leading to undesirable system behaviours. For operation and maintenance, AI created data may need computer-based interrogation due to size and complexity. However human oversight is required.

8 Dynamic hazards and safety arguments

Traditional top down and bottom up approaches are problematic for AI-based systems as AI power often relies on its emergent behaviours. AI is complex and difficult to deconstruct and approaches often fail to capture human-machine interactions. Techniques such as system theoretic process analysis that focus on system behaviour can address such challenges.

9 Maintenance and operation

Source data comes from operational, failure or adversarial domains. Determining the source allows the system to process or discard it and the identification of data drift. System maintenance can expose the confidentiality of its dataset, which may compromise system integrity.

10 Specification

A detailed safety requirements specification, produced at concept stage, will minimise rework, residual safety risks and provide a basis for validation. This is challenging as the complexity of AI system performance is difficult to analyse.

Conclusion

This is the first in a series of IET outputs on this topic. A more detailed document is currently being developed and will be published shortly.

All feedback on this paper is welcome. Please contact sep@theiet.org.
This paper has been produced by the Engineering Safety Policy Panel.
For more details on the Panel's work, visit theiet.org/engineering-safety.

@TheIET theiet.org

E6D22001/EngineeringSafety/0522