Seminar Report: Developing Safe AI Systems



All images © Alex King and Mike Parsons

The "Developing Safe AI Systems" seminar was held at the BCS in London on 27th June 2024. This seminar was all about how to develop and build safe AI systems: it discussed the methods, tools and techniques that should be employed, the all-important approach to testing and V&V, and the quality and safety assurance processes that govern what has to be done. The speakers outlined possible ideas and approaches for these evolving areas.

The five speakers presented insightful talks on developing safe AI systems, covering a range of topics:

- Richard Hawkins (University of York): Through-life assurance of ML systems
- Alan Simpson (Ebeni): Where are we with AI Safety?
- Iain Whiteside (ReSim): Testing Times Ahead for AI
- Gary Brown (Airbus): Certification Use Reliance (CURe) for an Aircraft Level view of an AI ML Part 21 integration
- Nick Tudor (D-RisQ): Robust AI Planning

There was also a discussion: do we have what it takes to develop Safe AI?

Richard Hawkins

Richard presented the Sense, Understand, Decide, Act (SUDA) model[1]. A question was asked about different driver behaviour in different countries. Richard replied that the fact the Machine Learning (ML) is safe in the UK doesn't mean it's safe in China. That's why requirements are so important.



Richard spoke of the importance of independent testing. Developers shouldn't know the verification dataset, otherwise they might train the system to pass the test. The testers should deliberately try and get the system to fail. The testing has to be realistic.

Richard also spoke about the importance of lifelong assurance. There can be unpredicted changes to the operating environment. How do we respond to maintain assurance?

Richard spoke about ML safety monitors. He cautioned that trying to characterise what an unsafe output looks like can be very difficult.

A question was asked about whether you could use completely simulated images. Richard replied that some people have given up using photorealistic images. Although the images look good to the human eye, different features are being activated in the ML. Self-driving car companies are using real-world images they've collected instead.

Alan Simpson



Interestingly, Alan's PowerPoint slides had been generated by AI. They were pretty good, but contained some hallucinations and a spelling mistake.

Alan presented the work of the SCSC Working Group on AI Safety, which leads on from the work of the Safety of Autonomous Systems Working Group.

Alan spoke about current research in AI for functional safety, including a plug for our friends at the University of York and their Body of Knowledge [1]. He also spoke about the IET report on The Application of Artificial Intelligence in Functional Safety [2] and the recently published ISO/IEC TR 5469:2024 on Artificial Intelligence – Functional safety and AI system [3].

It's difficult to keep up with the rapid changes in AI. Do we need more guidance, for example on how to deal with AI hallucinations? A key problem is that generative AI can make stuff up.

A survey by the Ada Lovelace Institute [4] found that the public doesn't see much benefit in self-driving cars and are concerned about them. A maxim from civil aviation is that you should only make a change if there is a notable improvement to operations or efficiency and no adverse effect on safety, or if there is an improvement to safety.

[1] Unfortunately, I missed the first half of Richard's presentation because my train, along with many others, was delayed due to a fatality between Reading and Paddington. Fortunately, Mike Parsons gave me a copy of his notes and I was also able to catch up with Richard's talk on the SCSC website.

Alan looked at how good is generative AI? A medical study of Large Language Model (LLM) accuracy in diagnosis found that ChatGPT 4 was 87% effective but was poor with rare or masquerading diseases.

Alan said there is increasing regulation and guidance on AI from the UK government and the EU (the AI Act).

Alan ended with a call for people to get involved in the Working Group, which would like to reach out to the wider safety community for concerns.

Iain Whiteside



Iain talked about embodied AI, which is AI that interacts with the world in some way in support of a task.

He made a distinction between the AI subsystem and downstream. Traditional ML metrics like precision/recall fade into insignificance. You need domain-specific metrics. He gave the example of 3D bounding boxes for autonomous vehicles, where larger errors are tolerable for objects further away.

I liked that he recognised the difference between verification and validation. The latter includes detecting contract breaches during real-world driving and downstream failures that occur without a contract breach.

In the early stages of full self-driving, it was easy to make progress. You didn't need much testing to find problems and improve. Almost every autonomous vehicle company took this route, but you grind to a halt. Enter video game designers. Teams started to build closed-loop simulators. However, different parts of the neural network light up when using simulators compared to the real world, even though the images look photorealistic. Waymo has driven over 2 billion miles in simulation, but self-driving continues to be just around the corner. The generative AI companies are making the same mistakes.

It's too much fun for robotics engineers to tinker with actual robots. There is a gap between systems engineering, robotics, ML and safety engineering. Software developers are interested in point-of-time tests. Safety engineers are interested in a more holistic approach. Systems engineers live in between.

Iain concluded that real world testing is precious but is also storage intensive. We need to be able to extract the most value from it.

Gary Brown



Aviation is a highly-regulated and safety-critical domain. Companies don't put ML on an aircraft if they don't need to – it's not a gimmick.

The certification basis remains CS-25. SAE ARP6983/EUROCAE ED-324 [5] is a draft Process Standard for Development and Certification/Approval of Aeronautical Safety-Related Products Implementing AI. ARP6983 is a consensus-based standard that provides a Learning Assurance Level (ML-DAL) with an associated set of objectives.

Gary talked about three pillars: Certification, Use, Reliance (CURe).

Nick Tudor



Nick gave a very interesting presentation on Robust AI planning. He described an autonomous vehicle that uses AI to allow real-time re(planning) on board the vehicle with no need for intervention by a human. This would particularly allow autonomous operation in environments with poor communications links such as underwater and inside buildings.

However, it's not possible to assure the behaviour of AI, so the system architectures 'cages' the AI with predictable behaviour. Formal proof of the safety monitor guarantees that the boundary cannot be violated and provides assurance behaviour. This enables the benefits of AI, while providing the necessary safety and security assurance.

The system is built using Planning Domain Definition Language (PDDL). Many general purpose solvers for PDDL-expressed problems exist, but these planners are constrained by the computational resources available on-board an autonomous system and they return plans in minutes or hours. D-RisQ will produce domain-specific real-time planners that are implemented using FPGAs, which allow parallel processing and are therefore massively efficient, allowing replanning in a few seconds. The PDDL is transformed into the hardware description language, VHDL, for implementation. A formal proof is conducted that the VHDL preserves the required safety properties of the PDDL.

References

[1] University of York, Centre for Assuring Autonomy, "Body of Knowledge", www.york.ac.uk/assuring-autonomy/about/aaip/body-of-knowledge
[2] IET (theiet.org), "The Application of Artificial Intelligence in Functional Safety", electrical.theiet.org/media/ifbjt25i/the-application-of-artificial-intelligence-in-functional-safety.pdf
[3] ISO, "Artificial intelligence – Functional safety and AI systems (ISO/IEC TR 5469:2024)", www.iso.org/standard/81283.html
[4] Ada Lovelace Institute, "How do people feel about AI?", www.adalovelaceinstitute.org/report/public-attitudes-ai
[5] SAE International, "Process Standard for Development and Certification/Approval of Aeronautical Safety-Related Products Implementing AI – ARP6983 (WIP)", www.sae.org/standards/content/arp6983

Report by Dewi Daniels

Dewi Daniels is an expert in the development and verification of safety-critical software. He has worked on many software projects across many industries. He is a Fellow of the British Computer Society. He was a member of the committee that wrote DO-178C and he is currently one of the experts representing the UK on the IEC 61508 committee.