Seminar: Deployment, Operations and Maintenance of Safe AI Systems



The "Deployment, Operations and Maintenance of Safe AI Systems" seminar was held at the Cumberland Hotel in London, on 26th September 2024 and hosted by Mike Parsons. The speakers presented a number of perspectives on bringing AI into service, including safe operation, adequate maintenance, human integration and organisational readiness.

The six speakers presented insightful and compelling angles on development and using frameworks for safe AI systems, covering a range of topics such as:

- Bob Oates and Carolina Sanchez (Cambridge Consultants): Organisational readiness for AI: Governance and Assurance – how we ensure that our organisations are ready and willing to integrate AI.
- Kate Preston (University of York): A human-centred assurance framework for deploying autonomous systems based on previous research and demonstrator projects – case studies and some early steps towards defining a human-centred assurance framework.
- Karin Rudolph (Collective Intelligence and the Ethical Technology Network): Responsible Deployment of AI Systems: What businesses need to know – the AI regulatory landscape, socio-technical systems, AI standards and the important societal impacts of AI.
- Jane Fenn (University of York): A New Approach to Creating Clear Operational Safety Arguments – how do we argue about operational safety and why is this not yet good enough, the complicating influence of AI.
- John McNicol (Nova Modus): Systems for Safety When Automated Vehicles Go Wrong! – AI in vehicles, new and emerging standards, the concept of minimal risk events and how considering remote operation safety can enhance AI vehicle systems.
- Richard Hillman (Horiba Mira): Safety assurance methods for automated vehicle deployments: learnings from the Harlander project – challenges of assuring automated vehicles and an introduction to the Harlander project.

Bob Oates and Carolina Sanchez



A double act from Bob Oates and Carolina Sanchez was a timely reminder of the fact that AI integration is not just about the technology: the organisations, stakeholders, individuals and human processes are also of paramount importance. Bob and Carolina asked us to consider the seven AI assurance pillars, and how AI amplifies existing safety challenges such as ensuring data integrity, adequate testing, and reliable software supply chains. The motivating example – a lawyer in the US using ChatGPT to perform legal research – resulted in a very lively response from the audience!

Bob and Carolina then moved on to discuss the topic of governance of AI systems. The emphasis here was very much on providing a grounding for this in best practice, and also on an examination of how organisational responsibilities for AI implementation don't always align with the way the technology itself is developed.

Bob and Carolina also challenged us to consider the processes and policies which organisations need to put into place to ensure responsible adoption of AI: a statement of ethics, AI usage policy and a system for AI risk management. The final takeaway was a very relevant observation here that we, as safety engineers, have to use assurance foresight so that we don't in future end up having to use the word hindsight: definitely a thought to chew on for the rest of the day!

Kate Preston



Kate Preston followed, with a whistle-stop tour of some of the existing research that considers human factors and AI ethics, including work from the Assuring Autonomy International Programme (AAIP), which the SCSC was lucky enough to hear about at a previous seminar. However, as Kate observed, many significant challenges still remain! Real-world AI integration has not yet made adequate use of the frameworks and guidance to ensure safe and ethical development, while it still remains unclear who takes responsibility for the AI.

Kate took us through some of the reasons for this, including the fact that the socio-technical aspects of AI are still being underconsidered: the message here is that the people are as important as the technology! A human-centred approach would apply human factors to ensure that the appropriate consideration is being given to how humans might interact with the AI-enabled systems, as well as how external systems and organisations might affect AI use.

Next, Kate introduced us to how such a human-centred approach might be developed, by a two-pronged attack consisting of a scoping review and a systematic analysis of existing research performed under the auspices of the AAIP. The initial findings for these were impressive; I think the extent of existing research surprised many of us, and provided an inspirational note on which to end Kate's presentation.

Karin Rudolph



Ethics and human factors remained firmly in the spotlight for Karin's presentation, looking specifically at responsible deployment of AI. Karin took us through the tangled web of legal instruments mentioned in the EU AI Act: a daunting reminder of just how much legal infrastructure is needed to begin to assure that AI is being adequately regulated. As Karin noted, we're currently in the "era of risks", and adequate AI governance is essential.

The topic of the environmental impact of AI came up here again, with many of us surprised to learn just how much worse AI is for the environment than flying, car manufacturing and fuel consumption! A timely reminder indeed that risks go beyond safety, and Karin emphasised this with a concise and helpful introduction to the National Institute of Standards and Technology (NIST) Framework for AI risk management. However, legislation is only as good as the thinking behind it, and Karin soon challenged us to don our thinking caps by asking for an exact definition of "ethics": we all know what we mean by it, but the importance of establishing a shared definition quickly became clear!

Karin also invited us to consider some of the fundamental socio-technical aspects of systems, with a thought-provoking example of how people can form emotional attachments to AI systems, and come to depend on these beyond their simple functional performance. The landscape of risk is wide ranging, and it was rapidly becoming clear that adequate management of all of these risks is necessary if we are to claim that we really are developing "responsible" AI.

Jane Fenn



After a break for lunch, Jane Fenn picked right back up on one of the major themes for the day: how do we ensure operational safety of AI systems? Safety has certainly changed in the past thirty years, and Jane took us through a number of significant safety events, from Piper Alpha to the Haddon-Cave report. It was rapidly becoming clear to all that operational safety has not always been considered with as much attention as it requires!

Jane highlighted some of the reasons for this, ranging from a lack of understanding of risk on the part of the operators, a general preference for "box-ticking" and simple rules rather than engineering judgement, and a lack of understanding of how the safety mitigations truly contribute to mitigations of hazards. A daunting task, but Jane emphasised that clear and explicit identification of operational safety requirements would go a long way towards addressing these. She proposed a separate – although linked – argument for operational safety, which has the advantage of hiding confusing design details while at the same time ensuring that the link between operational safety and design remains clear.

Of course, this applies to all systems, but Jane encouraged us to consider just how AI – as

a new and evolving technology – can enhance this challenge. It isn't always clear with an AI system exactly what needs to be monitored to provide assurance of operational safety, and to date there have been too many examples where operational safety is provided by simply handing control back to the human operator, regardless of how alert or situationally aware they may be!

John McNicol



John refocused our thoughts firmly on the automotive domain, with a presentation focused on automated vehicles and system safety. With mechanical failure, power failure and software failure all potentially in the offing, John introduced us to the new BSI Flex standard 1888 for considering just these eventualities. BSI Flex 1888 discusses both minimal risk manoeuvres and minimal risk conditions, and much of the material in this new standard has come out of the SafeMRX project exploring these.

John also reviewed the state of existing standards in this sector, noting in particular that these fail to adequately consider minimal risk manoeuvres or minimal risk conditions, in particular where busy roads or automated buses and trucks are involved. It was rapidly becoming clear that something new was definitely required, and John's presentation also offered a way for audience members to get involved via the industry consultation stage in November 2024.

This wasn't the only standard John introduced us to, though, with the presentation finishing up via a look at BSI Flex 1886, considering systems for remote operation safety of Autonomous Vehicles (AVs). Version 2 of this had just been published, and audience members very much appreciated the chance to walk through and see what was new in terms of remote operation systems, whether these be perception systems, communication systems or control systems. Coming to the end of a talk focused primarily on what could go wrong with AVs, the knowledge that these standards are out there felt definitely like a step in the right direction!

Richard Hillman



Richard continued the automotive focus with the final presentation of the day, an in-depth look at the Harlander project, and what we can learn from this about safety assurance methods for AV deployment. The Harlander project was new for some of the audience, who were intrigued to hear about the use of Belfast Harbour Estate for two autonomous vehicles, timed to link specifically with the rail schedule. Richard cautioned us about some of the ways in which the real-world setting can introduce challenges for AVs, however, with edge cases such as confusing marking and signage, and even the presence of animals on the road: surely the first time an SCSC presentation has considered runaway seals!

Moving on, Richard emphasised the need to consider best practice before executing any testing. This is particularly the case where AI is involved, as machine learning creates a large black box that means the traditional methods of cascading system level requirements to subsystems are made much harder. Looking further into specific forms of testing, Richard's presentation introduced the audience to three categories of scenario-based testing: functional, logical and concrete scenarios. While functional scenarios require simple description in natural language, concrete scenarios quickly require statistical sampling methods to help control test volume: with a reasonable real-world number of 30 parameters and 20 levels each, this generates $1x10^{39}$ test cases for a single logical scenario!

Having thoroughly convinced us of the need for intelligently-applied best practice, Richard then discussed how multi-pillar testing combines different forms of test evidence, but still is not sufficient to assure Safety of The Intended Functionality (SOTIF): rather, design analysis evidence is also needed. With SOTIF itself only being a tiny fraction of the overall safety case, this presentation had been a convincing way to motivate us to consider how safety assurance requires so much more than testing!

Mike Parsons – Wrap Up

Mike Parsons wrapped up the session by presenting the answer to a very timely question: why can't we simply ask ChatGPT to solve this for us? Having asked ChatGPT the query How can I safely deploy an AI system into service? Mike obtained what looked like a reasonable answer that focused on such aspects as design for safety, robustness, failures; testing and validation; scenario planning for failures etc. However, the audience were quick to realise that these statements hide some very difficult and complex challenges, and ChatGPT doesn't get us any closer to solving these!

At the end of a day focused on the challenges of AI deployment and operation, the audience were unanimous in agreeing on the necessity for human judgement, clear thinking and safety assurance that demonstrably considers the people, organisation and real-world challenges of AI systems. A thought-provoking day, and a great discussion to finish!

Report by Catherine Menon.

Catherine Menon is a principal lecturer at the University of Hertfordshire. Her research focuses on the intersection of trust, safety and ethics in public-facing autonomous systems. She has a PhD in category theory and has been involved in the development of regulatory standards and guidance for safety, security and ethics assurance within autonomous systems.