Seminar: Frameworks for Safe AI Systems



The "Frameworks for Safe AI Systems" seminar was held at the IET in London, on 25th April 2024 and hosted by Mike Parsons. The speakers examined the legal, ethical and regulatory frameworks around AI, as well as considering the social and societal implications of use of these systems.

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

- George Mason (Frazer-Nash): An Introduction to Safe AI – definitions, debates and comparisons between different AI techniques, plus the underrated importance of education
- Mark Sujan (Health Services Safety Investigations Body): Human-Centred Healthcare AI – putting the human back in AI, the tricky doctor-patient relationship and AI hype vs reality
- Zoe Porter (Assuring Autonomy International Programme): A Framework for the Equitable Safety of AI-based Systems – socio-technical systems, what happens when the risk-bearers aren't the beneficiaries, and the difficulties of reconciling ethics and regulation
- Dai Davies (Percy Crow Davis & Co): De-mystifying AI – the law and AI, limitations on practical uses of AI and the all-important conundrum of human equivalence
- Phil Mulvana (UKAEA): Design Frameworks for Autonomy – a strawman safety case using SACE, plus a real-world system that's a prime candidate for AI introduction
- Gabriel Pedroza (Ansys): Design, Exploration and Evaluation of Safety-Critical Software for Integrating AI-based Algorithm – a framework to address and explore AI safety, the three axes of autonomy, and the pillars to improve AI maturity

George Mason



George Mason got the day off with a bang by challenging us to consider whether we could all, in fact, define AI (stealing the thunder from our attempt when he confirmed that no single universally-accepted definition exists!) He talked us through the potential harm that could be caused by AI, and invited us to consider how recently safety has even been a key focus of AI at all.

Moving on to the legal side of things, George introduced us to the EU's AI Act, approved in February 2024, with a key priority to uphold human rights, this is one of the most significant advancements of AI regulation. The UK's approach, by contrast, aims to balance safety with innovation which is always going to be a tricky tipping point.

George walked us through the AI Standards Hub for a quick glimpse of the comprehensive database of AI standards, pointing out that most legislation is still under development or non-existent. The rate of AI development outpaces the rate of AI legislation development; this observation had some loud agreements from the lawyers in the room! To finish up, George emphasized the importance of education in terms of understanding AI: if we don't know how to use these systems safely and reliably, he asked, then how can we be sure of the potential harm they might be causing?

Mark Sujan

Mark Sujan was up next, with one key message for us all: language matters! He took us through the ways AI might be used in healthcare, from monitoring of disease outbreaks via social media (a reason to get back on Facebook?) to appraisal of literature to scheduling of operating-room time. We were intrigued to hear about some ChatGPT hallucinations that have popped up when it's asked medical questions; in some cases it has claimed to have a "masters degree in public health and personal experience with type 2 diabetes" (!!)



Next, Mark moved onto AI hype vs AI reality, noting that very few people are working on last-mile AI, i.e. the real-world implementation of AI into healthcare systems. One example of this would be using AI to help answer emergency calls. It was edifying – if unsettling – to brainstorm all the ways this interaction might go wrong with AI in the mix, but one question kept coming up again and again: will call handlers have the confidence and authority to override the AI when they think it's wrong?

Mark then talked us through the ways in which AI might change the relationship between a doctor and a patient, and queried the extent to which organisations are ready to adopt AI. Certainly many of them are keen to do this but – in an echo of the same concerns brought up earlier – there are distinct question marks over whether these organisations truly do have the AI literacy and capacity to manage this.

Zoe Porter

Zoe picked straight back up on the healthcare concerns, this time looking at them from an ethical perspective. She introduced us to the PRAISE framework (Principles-based Ethics Assurance, an inspired acronym!) for putting together ethical assurance cases. We learnt about the four+1 ethical principles of non-maleficence, beneficence, respect for autonomy, justice (+ transparency).



Regulation continued to be a hot topic here, as Zoe told us: regulation focuses on risk and safety, with very little around ethics of AI being legally required. The legal and regulatory constraints – or lack of them – of AI was certainly emerging as a common theme for the day! Zoe invited us to consider the difference between attempting to construct a top-down vs bottom-up ethics assurance case, this being the difference between assuring a single ethical property or a suite of ethical properties.

Zoe also took us through an example of the PRAISE framework: the DORA system is a conversational AI agent, used for follow-up calls after cataract surgery. I think we were all realising just how much conversation and dialogue is involved in the healthcare field by this point, and becoming increasingly well-informed of AI's ability (or lack thereof!) to carry out such conversations safely and ethically.

Dai Davis



Following a break for lunch, Dai Davis challenged us to think explicitly about what the law requires of AI, which was certainly a question on everyone's minds after the morning talks. He invited us to consider the training data for ChatGPT 4: this is in the public domain, and ended in September 2021. While augmented Large Language Models (LLMs) – being LLMs trained on specialised and updated text – can go some way to addressing concerns about these systems being out of date, the environmental issues remain a major problem. Dai illustrated this by showing us that it would take the amount of water required by the entirety of Chester for a month, just to run the servers! (Which would you rather have: Chester or ChatGPT?? Now there's an ethical conundrum…)

Dai moved on to talk about the limitations of AI and to consider what the human equivalent of machine learning would be: hyperthemesia, which for those of us new to the word, means a photographic memory. However, he cautioned us about trying to find true human equivalents: humans, he warned, can certainly censor information, but they do so with knowledge of what the true answer is. Systems like ChatGPT censor without having the full information, a much more dangerous set of circumstances.

Finally, Dai took us through some apposite – and fascinating – legal observations. We discussed Henthorn vs Fraser (withdrawing an offer to sell a house, after the acceptance of this offer was already popped into the post box); the advent of email would surely have made this kind of situation vanishingly unlikely. He also took us through the timescales for a law to go from proposal to implemented legislation (10 years, which surprised many of us!) and highlighted some of the ways in which the Copyright, Designs and Patents Act is unfit for purpose against AI.

Phill Mulvana

Phill picked back up on the theme of frameworks, introducing us to the SACE framework (Safety Assurance of autonomous systems in Complex Environment) for building an autonomous safety case. The application system was particularly interesting: the "gloveboxes" used in the nuclear domain to allow handling of items that are hazardous to humans. As Phill's slides showed, this work requires not only a highly skilled operator, but can be particularly uncomfortable for humans to do at all.



The high-level view of the project Phill presented emphasised input from key stakeholders to feed into the AI assurance case, as well as real hardware validation. There were some very interesting reasons why SACE is well-suited to construct the strawman, most notably that it is domain-agnostic. As Phill noted, this pre-empts any potential objection that you can't use – for example – a medical framework in a nuclear setting!

Phill gave us a few more behind-the-scenes observations, including noting that the team did not do the traditional safety and compliance process. They were given eleven months and did the first four planned phases:

1. Operating context assurance
2. AS hazardous scenarios ID
3. AS hazardous scenarios ID (again)
4. Safety requirements assurance.

He also described how some of the later stages might work, and finished with some great lessons learnt, including the very timely recommendation to assess whether AI is indeed truly essential for a given system!

Gabriel Pedroza

Last but certainly not least, Gabriel Pedroza took us through a methodology to design, explore and assess AI safety. Gabriel first invited us to consider the challenges for AI development, including the three axes of autonomy: environmental complexity, mission complexity and human independence. The ideal of an AI system is to mimic human physical, biological and social features, but as Gabriel described to us, the broad range of uncertainty makes this easier said than done!



Gabriel introduced us to a methodology for certification of AI, focusing on three standards or documents: ARP-6983, EASA Concept Paper V2, ISO PAS-8800. The fundamental point made here – as we've heard in previous talks – is that safety does not change in different domains. Quite a timely reminder here, in that it's very easy to become immersed in one domain and forget the transferability of safety principles!

Finally, Gabriel took us through safety assessment, using this methodology, of the UC-F16 Flight Formation. Some fantastic videos really showed the challenges of this and helped to give us a real idea of the context. The talk finished with a discussion of the pillars of AI maturity improvement: a common theme throughout the presentations has been the need to consider just how AI systems can be improved and matured!

Mike Parsons – Wrap Up

Mike Parsons finished the day by leading a general discussion, with the thought-provoking question: what would happen if you asked ChatGPT to suggest a framework for developing safe AI?? Mike had, in fact, done just that, and the results were surprisingly good. The problem – as quickly identified – is that using these kinds of systems will get you 90% of the way there, but unfortunately human nature is such that it's very easy to then stop thinking about the remaining 10%.

All in all, this was a very interesting and lively day, with some fantastic insights into AI frameworks and the challenges that must be overcome.

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 worked across industry and academia and has been involved in the development of regulatory standards and guidance for safety, security and ethics assurance within autonomous systems.