



**RAICo**  
ROBOTICS AND  
AI COLLABORATION

# APPLIED AI SAFETY SUMMIT

Findings Report

May 2024

# DOCUMENT INFORMATION

## AUTHORS

| Name                          | Title  | Affiliation                |
|-------------------------------|--|----------------------------|
| Phill Mulvana                 | Lead Technologist                                    | UK Atomic Energy Authority |
| Prof David Branson III        | Professor of Dynamics and Controls                   | University of Nottingham   |
| Dr Giovanna Martinez Arellano | Anne McLaren Senior Research Fellow in Industrial AI | University of Nottingham   |
| Dr Horia Alexandru Maior      | Transitional Assistant Professor in Computer Science | University of Nottingham   |
| Dr Jack C Chaplin             | Assistant Professor in Manufacturing Systems         | University of Nottingham   |
| Dr Pepita Barnard             | Research Fellow, Computer Science                    | University of Nottingham   |
| Dr Virginia Portillo          | Research Fellow, Computer Science                    | University of Nottingham   |

## PUBLICATION

First release: 9<sup>th</sup> May 2024

# TABLE OF CONTENTS

Document Information ..... 1

Foreword ..... 2

Workshop 1: AI Assurance ..... 3

Workshop 2: Responsible Innovation ..... 12

Workshop 3: AI Security ..... 21

Workshop 4: Panel Discussion and Wrap-Up ..... 29

# FOREWORD

Welcome. This report represents the culmination of our collective efforts at the RAICo Applied AI Safety Summit and the start of a shared movement towards a safer world made possible through the use of autonomous systems.

Across the following pages, we delve into content at the critical intersections of autonomous systems, safety, nuclear decommissioning and Fusion energy — areas that bring immense promise and a significant responsibility for those developing and operating it.

Throughout the summit, participants from diverse fields convened to address three pivotal areas: **security**, **responsible innovation**, and **assurance**, guided by experts from academia and industry.

This report captures the many thoughts were exchanged, insights gleaned, and learnings established. Synthesised here, it is my hope that it may serve as a foundation upon which we can stimulate further progress through collaboration and a shared understanding of challenges across domains.

I extend my deepest gratitude to all who contributed their insights, expertise, and dedication to this summit. Strengthening of the AS in safety critical environments is no easy task, but through the recognition of synergies between domains it is my belief that we can move towards a cleaner, more sustainable future.

**Phill Mulvana**



*Figure 1: Delegates of the workshop discussing responsible use of artificial intelligence.*

# WORKSHOP 1: AI ASSURANCE

## OBJECTIVE

The first topic of the AI Summit was that of AI Assurance. Although it is largely expected that future autonomous systems applications will be delivered with the support of AI, it is not clear yet which will be the appropriate techniques and approaches that will fully assure robustness and safety of these AI-based systems. This has created concerns about the risks and societal impacts associated with AI. Debates on the existential risks to humanity but also on more immediate concerns relating to risks such as bias, a loss of privacy and socio-economic impacts such as job losses has driven the current conversation about the future of AI systems. AI Assurance is consequently a crucial component of risk management frameworks for developing, procuring, and deploying AI systems (i.e. AI operationalization), as well as demonstrating compliance with existing and future regulation. Recently, the UK government published a white paper proposing an AI Assurance framework for measuring, evaluating, and communicating trustworthiness of AI systems based on 5 key principles<sup>1</sup>. However, as this is a general framework, it is important to understand what the current practice in industry is in terms of assurance, how the framework is understood and to what extent it can be applied to safety critical applications. In this context, the following objectives for this session were defined:

1. Identify how currently industries understand each framework principle and if it applies to their business.
2. Capture of current practice and approaches to AI Assurance
3. Identify the challenges in addressing the framework principles.
4. Establish main directions for further study with regards to AI Assurance in safety critical applications.

## KEYNOTE

Dr Greg Chance from Frazer-Nash introduced AI Assurance, some of the existing techniques for developing and measure trustworthiness of AI systems and key practices and the challenges when developing and maintaining these types of systems.

## WORKSHOP STRUCTURE

The AI Assurance was structured in three main parts. There was an introductory presentation by Dr Giovanna Martínez-Arellano, where she explained the UK Government Framework for developing Trustworthy AI systems, defining the 5 principles: **Robustness, Security and Safety, Fairness, Transparency and Explainability, Accountability** and **Contestability**<sup>1</sup>. She briefly explained how

---

<sup>1</sup> <https://www.gov.uk/government/publications/introduction-to-ai-assurance/introduction-to-ai-assurance#the-ai-assurance-toolkit>

these points affect both data as well as AI model operationalization and made a general introduction to AI Assurance Tools and what practice could look like with some use case examples.

After the introduction, delegates were separated into three smaller breakout rooms, and in each room tables of 4-5 participants were organized. The workshop was led by **Dr Giovanna Martínez-Arellano**, with the support of **Dr Horia Alexandru Maior**, **Dr Jack Chaplin** and professional facilitators from Frazer-Nash Consultants.

The session at the breakout rooms was then divided into three activities:

1. **Icebreaker** (10 minutes) – “AI Assurance going wrong” – each table was given 10 minutes to discuss a real case where AI Assurance has gone wrong and the consequences. Delegates were provided with three examples but were given freedom to discuss one of their own experiences, or that they knew about.
2. **Industrial Practice in AI Assurance** – Individual Challenges (15 minutes) – given the 5 principles, each participant will individually analyse how their business understands and to what extent addresses each principle using two perspectives: the data acquisition/management perspective and the AI model development /deployment perspective. Participants were provided a table and the following ‘scoring’ guidance as a starting point to fill in the table:
  - (1) Principle not addressed due to nature of the business or not understanding what it means to the business.
  - (2) Principle not addressed but desired/needed.
  - (3) Partially addressed (could be improved)
  - (4) Addressed (using tools and metrics)
  - (5) Addressed (with metrics and communicated to stakeholders)
3. **Table discussion and summary** (30 min) - from the individual activity participants were asked to share with the members of the table the results and come to an agreement on the following:
  - (1) Tables agreement of how each principle is addressed and understood.
  - (2) How each principle is measured

Each table chose a person on the table to summarise the discussions and present them to the room.

## KEY FINDINGS

The results of the activities in the workshop were captured, recorded, and analysed. Presented here are the key findings from the individual activity and what was captured during the table discussions. It is worth noting that participants came from wide range of industries, including academia, so a wide variety of opinions were captured. Findings are split into two the two analysis perspectives **Data Acquisition and Management** and **Model Development and Deployment**, and then splitting those to the 5 pillars.

# DATA ACQUISITION

In this section of the workshop, participants were looking at the different aspects with regards to how data is collected, annotated, documented, and managed in the context of their industry, the current practices and challenges around data for assuring systems that incorporate AI.

## SAFETY AND SECURITY

- There needs to be credibility and certification of where training data comes from and how it is used. But also processing (e.g. sampling information), cleaning etc. [multiple participants agreed]
- We're (nuclear) generally good at security, but it is often at odds with accountability. Sensitive data is secure but prevents its use in transparency of decisions. Can we provide open data sets for industry that can be used for training? [multiple participants agreed]
- Nuclear- it would be useful to have an agreed set of benchmark datasets. [one participant]
- We protect and anonymise data (score 4) but that makes it challenging when it comes to contestability [one participant]
- Data security and safety is partially addressed (score 3) [multiple participants agree]
- Data needs proper provenance, otherwise it is difficult to tell synthetic data from non-synthetic data. [one participant]

## ROBUSTNESS

- Redundancy in data collection. Use basic engineering principles rather than winging it. [multiple participants]
- Data needs to be continually gathered for both training and testing, as the use case environment keeps changing and adapting, otherwise initial test sets will have flaws. [one participant]
- Data quality – understand scientific approach for data collection, good design of experiments. This is desired but not the practice. [multiple participants]
- How do you spot systematic errors in training data? (e.g. dust on a lens). No current approaches for operationalization/monitoring/governance of the AI system development. [one participant]

## FAIRNESS

- Excluding some data for “fairness” can cause unintended consequences and secondary effects later.
- Can you perform formal analysis of fairness? There is a lack of understanding on the practices that apply to the assurance of fairness. [multiple participants]

- Difficult on context (decommissioning nuclear), but maybe “fairness” for technical data is not overfitting it? [multiple participants]
- Fairness among data sources – just because a sensor records at twice the frequency, does it have twice the influence on training data? [one participant]
- Multiple participants agreed fairness was not applicable or not understood how it would it be applicable to their context.

## TRANSPARENCY AND EXPLAINABILITY

- How is data processed? Not just what data is collected. There are no current practices that allow “correct” data governance, but also not an understanding what would “correct” mean for a particular scenario (multiple participants agreed)
- Can end users see training data to understand how it is used? How about the code? We don’t do this for “conventional” programmed systems. We trust current “conventional” systems to work, and don’t ask questions as a user (multiple participants agreed)
- Explainability appears to be the biggest technical hurdle in this area but there are likely to be many commercial and accountability issues around liability. Understanding whether the solutions for these should be technology-driven or policy/regulatory/commercial.
- Hard to be transparent in nuclear when there are measures of security (multiple participants agreed).

## ACCOUNTABILITY

- Most participants agree on a score of 2, where there is awareness of the need but not clear how it should be addressed, or practice is not in place.
- Can people be held accountable for providing incorrect training data? This is not currently done as there is generally poor practice with regards to governance (multiple participants agreed).
- Is it the machine holder’s duty to be accountable for data produced from a machine or the OEM? (one participant)
- There should be technical assurance certification for accountability of data. Employer to blame, risk assessment as tool to assess accountability of data (one participant).
- There should be clear records of how data was collected and why. Collection and processing must be well understood. This should be the norm, but it is not necessarily the practice (one participant).
- Companies point to raw data not filtered and processed, and so it is difficult to understand to what extent an issue may come from the data used to train a model (one participant).
- Data security can be misused to avoid taking accountability (one participant).
- Maintain clear records, who picked sensor or who made sensor (one participant).

## CONTESTABILITY

- “No obvious solution” (multiple participants agreed).
- What is contestability for data anyway? Not understood what it would mean for the context of the participant (one participant)
- For automation contestability may be counterproductive but this should be risk assessed. (multiple people agreed).
- There should be regulatory bodies responsible for contesting AI based systems (one participant).
- Effectively impossible? Timescales very short, AI systems require automated fail-safes (one participant).



*Figure 2: Dr Giovanna Martínez-Arellano (standing, left) listening in as delegates discuss the implications of assurance for AI deployment.*

## SUMMARY OF FINDINGS

From a data perspective, there is generally an awareness among participants about the importance of security measures when handling data, particularly sensitive data and in most cases, there are measures in place to handle security of data. At the same time, there is a general concern that this may act against transparency.



A strong link was identified by participants between the provenance, bias and robustness of data. With adequate documentation of how data is gathered (e.g. design of experiments, sensor details, data frequency), robustness may be achieved, provided both acquisition and provenance are a continuous practice. Provenance, however, is not currently a standard practice across different industries. For applications like nuclear, systemic bias from a social perspective was not identified, but mostly selection bias, where there is a potential of mistakenly leaving some variables or unseen scenarios out due to the complexity of the context.

With regards to explainability, it was recognised by participants that there is no proper documentation of how data is pre-processed and prepared (e.g. handling missing values, data splits, balancing data, feature extraction approaches) for the application of AI techniques. This has been acknowledged by the participants to be key for achieving scalability of AI in industry and being strongly linked to accountability as well as robustness.

Accountability and contestability were interesting topics of discussion both in terms of data and model development and provenance as an underpinning factor for achieving both. As with other principles, there is an agreement that there should be accountability as part of the assurance framework, however, it is unclear which actor is accountable for each of the multiple tasks that are part of the whole AI system development cycle starting with data acquisition. There were questions regarding accountability of the data sources. Who is made accountable? The equipment manufacturer, the person in charge of deploying the sensor, the person in charge of managing the acquisition and storage of it, or the person who pre-processes it for AI development? And how do we contest data when provenance is not addressed? These are questions that need to be addressed by sector.

## **AI MODEL DEVELOPMENT AND DEPLOYMENT**

### **ROBUSTNESS**

- A general agreement from participants that the principal is understood however poorly addressed as it might not be entirely clear what the “best” approaches to ensure robustness would be.
- Companies don’t understand how hard it is to turn data into a reliable AI system – end up with something unreliable (multiple participants agreed).
- Most people are taking existing AI systems they don’t understand and using them. But what if people try and modify them? How can they be guided to make sure it remains robust? (multiple participants agreed)
- Lighting conditions known to negatively impact machine vision systems. What are the “lighting conditions” for other AI applications?
- Depends on architecture, but there should be frameworks to allow robust deployment.
- Extensive testing across multiple scenarios, partially addressed, but they still would not work for all situations. (multiple participants agreed)

- Academic work doesn't emphasise robustness often applying 3rd party technology, only validating within specific context.
- AI requires an adaption of the hierarchy of control we use elsewhere (multiple participants agreed).
- There is no need to reinvent the wheel, can we validate and verify AI systems the way we do "conventional" systems with existing standards? (multiple participants agreed).

## SECURITY AND SAFETY

- Multiple participants did not address particularly the security aspect of AI models.
- AI life cycle of expected (and unexpected) use should be considered by the user of the system.
- Can we measure against humans, using the same methods we use to measure humans? Is "as good as a human" a good metric? Humans also fallible (multiple participants agreed).
- Impact of noise may pose a threat to safety.
- Same AI solution used for different applications incorrectly. This may be a combination of being a black box as well as lack of understanding of the applicability and limitations of the technology.
- Consider whole lifecycle and environment. Appropriate measures or techniques, consistency, at least as good as human (one participant).

## FAIRNESS

- According to several of the participants answers, fairness is either not well understood or applies to their context (scores 1 and 2).
- Checks on unintended diverse discrimination needs to be addressed by the development team. Secondary effect of development decisions can be addressed using techniques such as latent factors. (one person).
- Poor ethics is the natural outcome from poor assurance (multiple participants agreed).
- For several nuclear AI applications, fairness didn't seem to be a principle that was relevant for assuring AI systems.

## TRANSPARENCY & EXPLAINABILITY

- There were multiple participants that understand the value of transparency and explainability, however very little is done to address these principles (score 2)
- Can we have open code and data? Even if we did, would anyone look?
- Quality of a model is not the same as assurance.
- Many companies circumvent transparency rules to remain a competitive edge.

- We don't report failures in AI development and design mostly because of the data burden - this should be captured to understand risks of it when it goes wrong, and its potential biases.
- Current explainability techniques don't show why it fails, just where the system failed. Root cause analysis is needed to improve the system. Just making code open does not allow transparency or to improve safety [multiple participants agreed].
- We do get visual linguistic explainability to understand why a model came to a specific output (score 4) [one participant].
- Maybe impossible and possibly not necessary? If we trust "conventional" systems without knowing how they work, why do differently with AI? [multiple participants agreed].
- For system assurance, we have to be transparent to HSE with documentation (score 4)

## ACCOUNTABILITY

- AI companies trying to pass liability from developer to user. (sometimes opposite too) [one participant].
- All AI accident court cases ended with settlement – a court ruling would set a precedent. [one participant]
- Nuclear very risk adverse, accountability may be more complex?
- Anyone can write (bad) code. Does that make them liable? [one participant]
- There are standards for programming safety critical systems. Where is this for AI? [one participant]
- Government wants a council to accredit Cybersecurity. Can we have one for AI?
- Who is liable and responsible? The user? But are they informed? Passing responsibility user + developer, there should be clear lines. [multiple participants agreed]

## CONTESTABILITY

- Divided opinions on contestability, as in some contexts it seems to be vital to the assurance of robustness although it might not be current practice, while in some contexts might hinder automation.
- With rule-based systems you can access the rules. AI should explain itself [one participant].
- Empower operators to always be able to reject a system's output [multiple participants]
- How do you prove systematic failure? How are several "one off accidents" shown to be related?
- Contestability may hinder automation (multiple participants agreed)
- Effectively impossible? Timescales very short, require automated fail-safes [multiple participants].

- AI in design engineering should be contested.
- Test in court. Audit trail to understand decision. Access to rules if rule-based system
- Certification systems for safety critical systems, challenge- regulation, charter for AI engineers

## SUMMARY OF FINDINGS

In general, across all principles, although some apply and some don't according to participants opinion, answers reflected a lack of current practice to address and measure them when talking about AI model development and deployment. In terms of robustness, the data reveals a lack of understanding of limitations of different techniques as people use models as "off the shelf" technology. AI users need to be *educated* about the key aspects of AI operationalisation, and how continuous model monitoring and redeployment is critical for ensuring robustness. Safety came as a central point for nuclear applications for which multiple participants proposed the use of fail-safes and how AI systems could be slowly introduced, and not take full control until stability of the AI system is understood or achieved.

Security was an interesting topic with regards to AI development and deployment. It was perceived that although there is awareness and current practices in place for data security, this is not the case for security during model development and deployment. This is another result that highlights the need to educating the public on software practices and standards, but at the same time, it highlights the importance of further understanding to what extent current standards are really applicable and in which case more work is needed for developing new standards.

Transparency doesn't seem to be valuable from an AI model point of view if it is not accompanied with explainability as well as provenance of the decisions made during development. These are things that need to be addressed together to deliver value and enable model robustness. Accountability, although understood and needed across all industries, it's the principle which participants were most unclear about what is the "correct" way of addressing it, and who is accountable for what. The AI life cycle is a long, complex and continuous process for which multiple decisions are made, making it very challenging. Finally, contestability was an interesting principle as it is perceived as a potential for hindering automation. It was highlighted on multiple occasions that we trust "conventional" systems despite not necessarily knowing as a user how they work. Why can't we take the same approach with AI systems? The difference, however, is that although from a user perspective we don't know how a conventional system work, that is not true for the system developer. The system developer would go through thorough validation and certification procedures and would be certain about the expected behaviour of such a system. With AI systems, in the contrary, unless they are designed with fail-safes and constrained, behaviour may be unexpected. With the increasing concerns of the potential negative impacts of AI misuse, there is a need for sharing best practices and more work on the development of standards for ensuring safety and robustness of AI systems.

# WORKSHOP 2: RESPONSIBLE INNOVATION

## OBJECTIVE

The key topic covered in the workshop was Responsible Innovation (RI), a continuous process to anticipate how research and innovation outcomes and processes may affect people and the environment in the future, and act in the present to gain most benefit, minimise risks and avoid harm. There remains a gap between theory and practice of RI, so this workshop's objective was to address this challenging gap through a facilitated, interdisciplinary collaborative activity using [Responsible Innovation \(RI\) Prompts and Practice Cards](#), engaging participants in discussions related to specific examples within Autonomous Systems, Artificial Intelligence (AI), and Robotics in the nuclear sector. The summit offered an opportunity for a hands-on experience involving a diverse group of participants, including key stakeholders, to systematically identify responsibility challenges, reflect on practices, adopt approaches and solutions from other sectors, and create action plans to ensure inclusive practices, foster ethical and responsible decision-making, and embed RI into practice.

### **What is Responsible Innovation (RI)?**

RI is an iterative and evolving process by which societal actors and innovators become mutually responsive to each other with a view on the ethical acceptability, sustainability and societal desirability of the innovation process and its marketable products in order to allow a proper embedding of scientific and technological advances in our society<sup>2</sup>. Alternatively, RI is doing research and innovation in a way that anticipates how it may affect people and the environment in the future and acting in the present to gain the most benefit, minimise risks, and avoid harm<sup>3</sup>.

This session's objectives were to:

1. Introduce tools to support RI practice.
2. Conduct a facilitated hands-on workshop focused on understanding key ethical and responsibility challenges/opportunities applied to AI powered systems in the nuclear sector.
3. Determine and discuss possible first steps / solutions / priorities to these problems (e.g. knowledge transfer opportunities and lessons learned from other sectors).

---

<sup>2</sup> Von Schomberg, Rene. "Towards responsible research and innovation in the information and communication technologies and security technologies fields." *Available at SSRN 2436399* (2011).

<sup>3</sup> Greenhalgh, Chris. "Responsible Innovation (RI) Prompts and Practice Cards (version 3.1. 1, November 2023)." (2023).

The workshop team was led by **Dr Horia Alexandru Maior, Dr Pepita Barnard and Dr Virginia Portillo**, a team with extensive expertise embedding RI practice and RI training, and co-facilitated by **Dr Jack C Chaplin, Dr Giovanna Martinez-Arellano** and professional facilitators from Frazer-Nash Consultants.

## KEYNOTE

Prof. Michel Valstar, co-founder of BlueSkeye AI and an Honorary Professor in Automatic Human Behaviour Understanding in Computer Science at the University of Nottingham, gave the keynote speech. Michel is an expert in machine learning, computer vision and facial expression analysis, and a recognised leader in affective computing and social signal processing. With over 18 years in this field, his work has been cited over 18,000 times in over 100 peer reviewed publications. His talk showcased BlueSkeye AI's perspective on adopting responsible AI technology in clinically trusted applications they have developed, and how RI is embedded in each development step.

## WORKSHOP STRUCTURE

The RI workshop took place at the end of Day One and involved a group of over 40 participants and stakeholders from across sectors. The workshop lasted approximately 2 hours and the session was divided into the following parts:

- **Introduction to RI** presented by Dr Horia A. Maior, Dr Pepita Barnard and Dr Virginia Portillo (20 minutes – all participants in the main room)
  - We provided definitions and examples of RI, including the AREA 4Ps- Framework for RI<sup>4</sup>, and introduced the UKRI funded networks Responsible AI (**RAI.co.uk**) and Trustworthy Autonomous Systems (**TAS.ac.uk**).
  - Introduction to the [Responsible Innovation \(RI\) Prompts and Practice Cards](#), a practical toolset designed to prompt reflection and discussions concerning responsible considerations as part of every step of research and innovation processes. This toolset was subsequently used by the participants in the hands-on, facilitated activity as described below.

At this point participants were split into three breakout rooms to discuss the following themes:

- Room 1: Public perceptions of nuclear and autonomous systems
- Room 2: Workforce related challenges for autonomous systems in the nuclear context
- Room 3: Robotics and autonomous systems in nuclear decommissioning

Each room was equipped with two open packs of RI cards, two A1-sized posters, printed with the relevant room theme name and headings, laid out on tables to structure the

---

<sup>4</sup> Jirotko, M., Grimpe, B., Stahl, B., Eden, G., & Hartswood, M. (2017). Responsible research and innovation in the digital age. *Communications of the ACM*, 60(5), 62-68.


discussion, with paper, pens, markers and post it notes to capture the key points as per Figure 3 and Figure 4.



*Figure 3: Thematic room set up for the RI workshop. Notice the Responsible Innovation Prompts and Practice Cards, a poster, pens, and markers and post-it notes. One facilitator was allocated to each table, giving everyone voice for their input and identifying the key take aways discussed at the table. All participants were engaged in providing their insights using the post it notes and poster template provided.*

- **Hands on activity** (60 min – participants split into three thematic rooms)  
In collaboration with RAICo and the University of Nottingham team, three **key themes** were proposed as contexts for the RI workshop:
  1. **Public perception of nuclear and autonomous systems,**
  2. **Workforce related challenges with autonomous systems in the nuclear context,**  
and
  3. **Robotics and autonomous systems in nuclear decommissioning.**
- Participants joined tables of 6-7 members and worked collaboratively approaching one of the above themes, according to their allocated room and were instructed to follow the following steps:
  - *Identify a named individual to present an overview to the other groups at the end.*
  - *Familiarise themselves with the RI Cards*
  - *Collaboratively propose a set of possible scenarios or identify areas of concern in line with the room theme.*

- Fill in the table (Figure 4) with pens and post it notes.
- In the last part prepare an overview of the 3 takeaways from your table to be reported back to the main room

|  <b>RAICo</b><br>ROBOTICS AND<br>AI COLLABORATION |                          | <b>Workshop: Responsible Innovation and Ethics</b><br><b>Theme name: Public perception of nuclear and autonomous systems</b> |  |        |                     |
|--|--------------------------|--|--|--------|---------------------|
| Context discussed  | Benefits / Opportunities | Issues / challenges  | Possible impacts (short and long term) | Action | Timescale/ Priority |
|  |                          |  |  |        |                     |

*Figure 4: Example of an A1 poster provided at each table in the RI workshop. During the 60 min exercise, participants used the RI Cards to raise issues as well as opportunities related to ethical and responsible development of robotics, and AI systems in the nuclear context.*

- **Summary** (20 minutes). The facilitator together with the other participants, prepared a short summary of the main findings and moved to the main room.  
At this stage participants all returned to the main room.
- **Presentation of key findings** (20 min). One representative from each of the groups in the three rooms prepared a short summary with key findings, challenges, and possible solutions to all summit participants. This was an excellent moment for reflection before closing the day.

## KEY FINDINGS

Presented here are the key findings from the activities in the RI workshops. Key findings and take-aways would sometimes overlap across the three themes.

## PUBLIC PERCEPTION OF AUTONOMOUS SYSTEMS (AS) FOR NUCLEAR APPLICATIONS

Issues discussed in this group included the public fear of AI, the lack of understanding surrounding this area, unequal access (social or geographical). Possible solutions included the need to do more research to better understand public perceptions towards AS and the nuclear sector. Also, the need to better engage with different stakeholders including the Government to get support on promoting more education in this area is key. Some concerns were related to changing public acceptability of risk in the nuclear sector, which might have a significant impact. Some of the output can be seen in Figure 5 and Figure 6.





## BUILDING TRUST

- Deploy in familiar sector.
- Tuned Messaging.
- Involvement of the Regulator (across different fields).

## ENGAGEMENT

- Research outreach and public engagement.
- Stakeholder input, Public Dialogue,
- Explainability of AI and AS
- Education, Re-skilling, Up-skilling, Multidisciplinary.
- Benefit to Society – defining and measuring benefit and disbenefit of AI.

## SUPPORT

- Financial Aid to SMEs
- More research on public perception (multiple demographics) is suggested by multiple groups!
- Community Building through similar events and summits
- Building clusters of expertise/Knowledge Exchange

## WORKFORCE RELATED CHALLENGES WITH AUTONOMOUS SYSTEMS IN THE NUCLEAR SECTOR

Some of the key highlights (also presented in Figure 7) are:

- Engage the workforce from the beginning
- Human AI Teaming brings new challenges
- Workforce Readiness: Upskilling / Reskilling existing sectors
- Educational opportunities
- Concerns re. De-skilling workforce
- Perception of AI in job satisfaction
- Identify unanticipated outcomes:
  - Clarity around accountability. As systems become more complex, accountability can become intimidating, potentially deterring professionals from safety-critical sectors.
  - Structures to inform workers.
  - Design to manage risks.
  - Follow best practices ([rai.co.uk](http://rai.co.uk))
- Justifying why AI vs Humans

- AI exacerbating inequality
- Objective decisions, control for bias.
- Regulatory Challenges: Rapid technological advancements challenge regulatory bodies to maintain a balanced approach to governance. Workers are often affected by this.

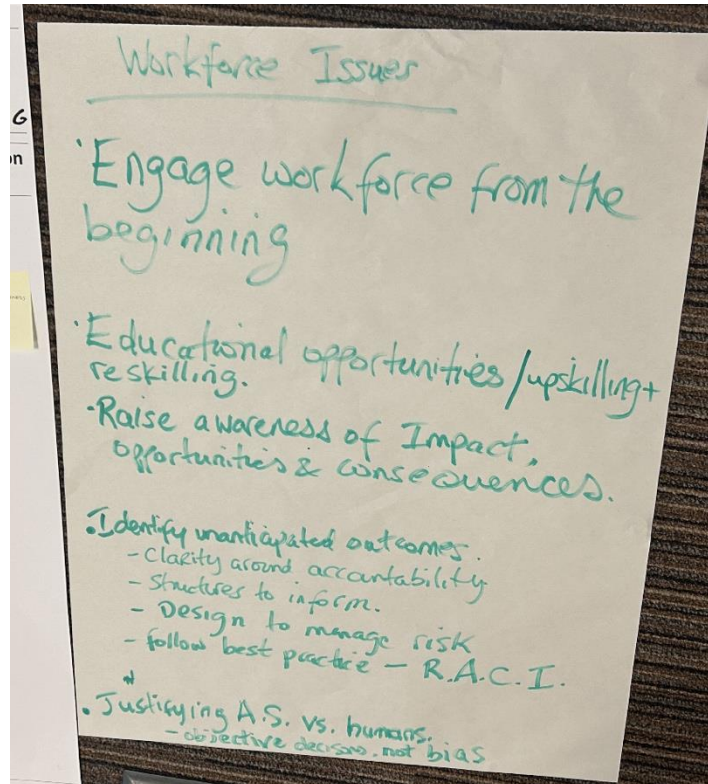


Figure 7: Outputs from the themed discussion on workforce related challenges with autonomous systems in the nuclear sector.

## ROBOTICS AND AUTONOMOUS SYSTEMS IN NUCLEAR DECOMMISSIONING

Some of the key contexts discussed in the room included the use of a SPOT-like robot for various maintenance and clean-up of nuclear waste, Computer Vision for damage recognition, drone inspections, Large Language Models to record queries, Advisory systems for waste sorting, as well as Robotics to replace human operators in hazardous environments. Some of the key challenges/solutions identified and discussed are (also presented in Figure 8).

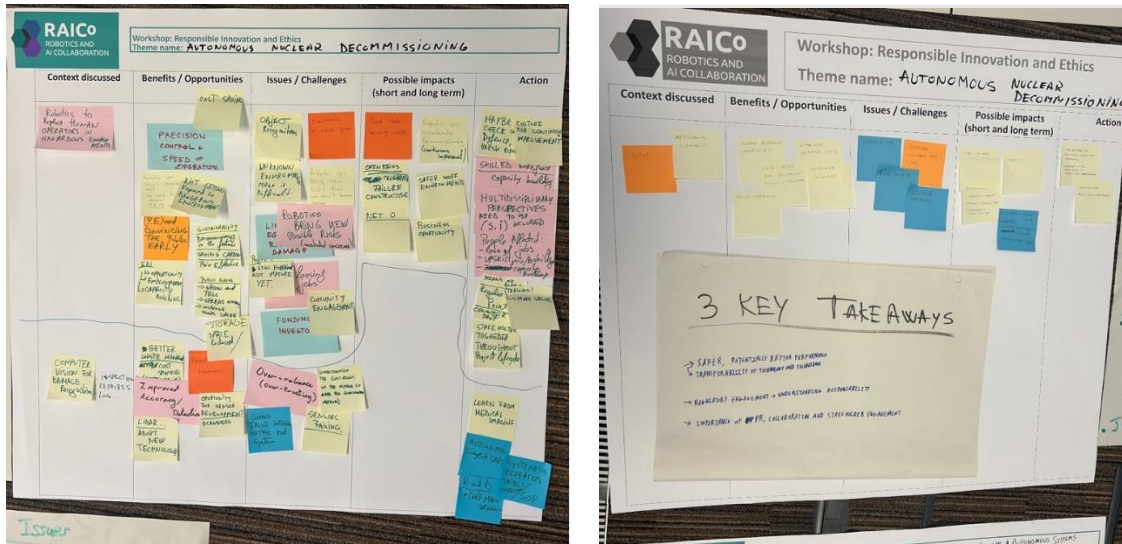


Figure 8: Responsible Innovation (RI) workshop focused on exploring implications in autonomous nuclear decommissioning.

3 Key take aways are also highlighted.

- Common Concerns across different sectors
  - Overreliance of Autonomy can have high consequences in the nuclear sector, depending on the application.
  - Failure/Sensors management. Accountability?
  - Who are the affected people?
- Lessons from other sectors
  - Computer vision problems discussed have already got solutions in healthcare (medical imaging) and manufacturing (chip inspection), a source of knowledge transfer.
  - Training, Upskilling, Capacity building
  - Managing expectations (short- and long-term impact)
  - Communication is important. Build Trustworthiness.
- Stakeholder engagement.
  - Regulatory engagement, Understanding Responsibility.
  - Collaborations across Industry, Academia, and Regulatory bodies
  - How do you engage the right people and where do you find the pools of experts you need?
- Cost effectiveness while improving people’s safety.

### SUMMARY OF FINDINGS:

We have identified the following actions needed shared between the three breakout rooms:

- More opportunities for stakeholder engagement (multi- and interdisciplinary),
- More education/skills needed,

- Better communication strategies at all levels from the research, industry and regulatory sector to better tackle the different challenges discussed withing each theme.

## OVERVIEW AND FUTURE IMPLICATIONS

- The Responsible Innovation Prompts and Practice Cards used in the workshop were welcomed and adopted by different participants including regulators as well as different UKAEA groups. Some attendees requested physical decks of RI Cards to use within their current projects at their institutions.
- New collaborations were established, and follow-up activities were planed for after the UKAEA AI Summit.
- Many participants highlighted the need for similar workshops and events in the future.



*Figure 9: Delegates from a breakout group reporting their findings and priorities back to the wider consortium.*

# WORKSHOP 3: AI SECURITY

## OBJECTIVE

The third key topic covered in the workshop was that of **Security and Safety for Autonomous Systems**. AI and autonomous systems have huge potential to change the way the nuclear sector (and others) operate, but it carries with it risks – both in terms of new tools potential hacks could have to assail systems, but also new vulnerabilities introduced by these new technologies.

Four objectives were defined for this session:

1. Understand the key challenges in securing autonomous and AI-driven systems, and in making them safe.
2. Determine the solutions to these problems (in an ideal world).
3. Elicit the key barriers preventing us from implementing these solutions.
4. Identifying any opportunities or accelerators which are underleveraged.

By identifying the barriers and underleveraged opportunities, early progress can be made towards accelerating the adoption of AI while remaining safe and secure. It was emphasized to delegates that a) the focus here was on AI and autonomous systems, and b) that any concrete experience is more valuable than hypotheticals.

## KEYNOTE

A keynote presentation was given by Dr Tarek Gaber, Senior Lecturer in Cybersecurity at the University of Salford. The talk primed the delegates on some potential safety issues and attack methods on AI methods, including issues around bias and fairness (tying into the previous day's session on ethics), privacy concerns, security risks, and regularity and legal challenges.

## WORKSHOP STRUCTURE

Taking place at the start of day two, the safety workshop took place over the same three rooms with (approximately) the same groups as the sessions on day one. Not all delegates were able to make both days, so there was some variation in group composition. The workshop was led by **Dr Jack C Chaplin**, with the support of **Dr Horia Alexandru Maior**, **Dr Giovanna Martinez-Arellano** and professional facilitators from Frazer-Nash Consultants.

The safety workshop session was divided into four activities:

- Icebreaker (10 minutes) – **What Does Security Impact?** Although groups were largely the same as the previous day, some changes had taken place. This ice breaker was an opportunity for groups to get to know each other, and brainstorm domains in which security has an impact. The content generated from the icebreaker was also used as an input to the challenges activity.

- Security Risks and Challenges (25 minutes) - **Key Security Risks and Challenges for Autonomous Systems.** Building on the ideas generated in the icebreaker, this activity asked delegates to think of ways in which autonomous systems could be attacked or disrupted, as well as challenges that autonomous systems present in terms of their security. These were then mapped onto two axes, one for the impact of the risk or challenge, and one for the likelihood or risk or the difficulty of the challenge. This allows for some degree of prioritization, which high impact / high likelihood risks needing more consideration than low impact / low likelihood.
- Solutions to Risks and Challenges (20 minutes) – **In a Perfect World, How Would you Solve your High Priority Risks and Challenges?** Focusing on the high priority risks and challenges identified in the previous activity, delegates were tasked to propose solutions to those issues. Delegates were asked to propose solutions with no restrictions around cost, technical hurdles, regulation, or restrictions. The solutions were then mapped against two axes, one for the benefit or reward of implementing the solution, and one for the cost and effort of the solution.
- What’s Stopping Us? (30 minutes) - **Why can’t we implement these solutions now? and What opportunities exist which need leveraging better?** The final activity focuses on the idealized solutions from the previous activity and asks – if we already know the solutions to these problems, why haven’t they already been solved? Delegates were asked to look for barriers inhibiting the solutions, but also to identify any opportunities that exist which should be leveraged better.

## KEY FINDINGS

The results of the activities in the workshop were captured, recorded, and analysed. Presented here are the key findings from the activities on Risks, Solutions, Barriers, and Opportunities.

All the activities had responses which could generally be categorised into four themes:

- **Technical** – barriers and opportunities that relate to the technical aspects of AI and autonomous systems, their implementation, performance, and usage.
- **Financial** – barriers and opportunities that are related to direct financial cost.
- **People and Skills** – barriers and opportunities that relate to people and their behaviour, and the absence or lack of relevant skills and training.
- **Regulation and Politics** – barriers and opportunities relating to regulation and governance, as well as potential political aspects of the technologies.

## SECURITY RISKS AND CHALLENGES

The risks and challenges activity allowed participants to prioritise risks against impact and likelihood. The responses presented here are those which were positioned in the high likelihood / high impact quadrant (i.e., the most severe challenges), and only responses which were shared by at least two groups are included here – although only one of the responses needed to be classified

as high likelihood / high impact to qualify (not all groups agreed on the severity of problems). Responses are also filtered to those which relate to security and safety of autonomous systems, rather than AI in a wider context.

## TECHNICAL

- The use of AI or autonomous systems for safety applications may make them vulnerable to new attacks, and the impact of these could be the most severe [3 responses HL/HI, 0 others].
  - Potential for non-secure data about safety systems to be used to compromise them, and concerns about compromised autonomous safety systems being harder to detect than non-autonomous ones.
- AI systems which rely on large, public data sets for training could be vulnerable to poisoning of those training sets [2 responses HL/HI, 3 others].
- Attacks on sensors could cause autonomous systems to behave in unexpected or deliberately malicious ways [2 responses HL/HI, 2 others].
  - Potential attacks include jamming sensors (LIDAR highlighted here), spoofing sensor readings, and undermining the prioritisation of sensor data.
- Attacks on networking infrastructure could cause autonomous systems to degrade or fail [1 response HL/HI, 1 other].

## FINANCIAL

- No high likelihood / high impact financial risks or challenges were identified.

## PEOPLE AND SKILLS

- Reduced human attention and oversight could make it harder to detect compromised systems, and for safety-critical systems, are autonomous systems more reliable than a person? [2 responses HL/HI, 0 others]
- Compromised autonomous systems could be used to harm people, with criminality, terror, political statements, or corporate espionage as motivation [1 response HL/HI, 4 others].

## REGULATION AND POLITICS

- No high likelihood / high impact financial risks or challenges were identified.

## SOLUTIONS TO RISKS AND CHALLENGES

The solutions activity also allowed participants to prioritise potential solutions to high-impact risks against two axes, expected cost, and expected benefit. The responses presented here are those which were positioned in the high impact and low-cost quadrant half (i.e. quick wins). One additional point was raised by many groups in the high-impact and high-cost quadrant which is



also reported. Only responses which were shared by at least two groups are included here – although not all groups agreed on the impact of solutions, only one needed to consider it to be high-impact to be included here. Responses are also filtered to those which relate to security and safety of autonomous systems, rather than AI in a wider context.

## QUICK WINS (i.e. high impact, low cost)

### TECHNICAL

- Regular security drills and use of red teams, updated to include attacks on AI systems *[2 responses HI/LC, 1 other]*.
- Keep AI segregated to control systems, and away from safety systems *[2 responses HI/LC, 1 other]*.
- Use of AI to automated security and resilience testing, offering a more frequent and cost-effective tool alongside red teams *[2 responses HI/LC, 0 others]*.
  - One example given for resilience testing is Netflix's Chaos Monkey, which autonomously terminates services and processes in Netflix's servers, to test the system's fail safes and resiliency<sup>5</sup>.
- Ensure that standard cybersecurity principles like least privilege and zero trust are applied to AI and autonomous processes too, to limit their influence *[2 responses HI/LC, 0 others]*.
- Air gapping of AI systems from critical systems on the network, ensuring the AI cannot influence other systems, even if compromised *[1 response HI/LC, 2 others]*.

### FINANCIAL

- No financial quick wins were identified.

### PEOPLE AND SKILLS

- Implementing continuous cybersecurity training for staff, to ensure the rapidly evolving threats (particularly around AI) are communicated *[1 response HI/LC, 4 others]*.
- Fostering a security culture that promotes transparency and support around security risks, in the same way that safety is considered now *[1 response HI/LC, 2 others]*.
  - Given the potential for damage and injury posed by compromised industrial systems (autonomous or not), security risks *are* safety risks.
- Collaboration between sectors and disciplines, to ensure threats and opportunities are shared and understood *[2 responses HI/LC, 0 others]*.

---

<sup>5</sup> <https://netflix.github.io/chaosmonkey/>

## REGULATION AND POLITICS

- Establish what good and responsible AI design and assurance looks like, with certification for well-managed systems [2 responses HI/LC, 1 other].
- Establishing standards and regulations around the responsible use of AI and autonomous systems, but also enforce these [2 responses HI/LC, 0 others].
- Creation of security-by-design standards and best practice for AI and autonomous systems [2 responses HI/LC, 0 others].

OTHER IMPORTANT SOLUTIONS (*i.e. high impact, but high cost*).

## TECHNICAL

- Using AI surveillance to detect attacks and breaches, and potentially stop them [3 responses HI/HC, 0 others].



*Figure 10: Workshop delegates in the red breakout group discussing the security implications of AI.*

# WHAT'S STOPPING US?

Whereas there was less commonality between groups for the risks and solutions activities, the activities on barriers and opportunities showed significantly more alignment and agreement between groups.

## BARRIERS

The responses to the barriers exercise are ranked by the number of times a specific barrier was mentioned, or where several barriers can be categorised together by the number of responses in that category. Responses were also filtered to keep only those that relate to security and safety of autonomous systems, rather than AI in a wider context, and only responses that were shared by at least two groups are included here.

### TECHNICAL

- The rate of change of technology makes developing security solutions difficult [2 instances].

### FINANCIAL

- There is low funding (public and/or private) for investment in security compared to other major economies [6 instances].

### SKILLS AND PEOPLE

- Taking security for granted, a lack of prioritisation for security, overconfidence, complacency, and a philosophy of “it will never happen to us” [7 instances].
- There are significant issues inhibiting the adoption of security and security practise [5 instances]
  - Five specific issues which were listed across several responses include fear of failure, resistance to change, bureaucracy, laziness, and cultural obstacles to innovation.
- A key skills gap for AI and autonomous systems, and for security individually, compounded by considering the two together [5 instances].
  - Specific skills-gap related concerns include a lack of education on the topics [2 instances], the multi-disciplinary nature of the problem, the lack of existing skills in these areas within organisations, and potential age biases when it comes to reskilling employees.
- A lack of communication between stakeholders and departments within an organisation [2 instances].
- A lack of awareness of the security risks of AI [2 instances].
- A lack of cooperation between industry and academic institutions [2 instances].

- Specifically, industry and academia not understanding each other's problems, and industry often being unwilling to share problems with academia.

## REGULATION AND POLITICS

- The challenge of creating international regulation on AI and its applications [*4 instances*].
  - Concerns that international regulation may not be specific enough, and that there will be differences in what is considered ethical and permissible between different countries.
  - Also concerns around who is trusted and credible enough to lead international regulation and can foster collaboration between non-friendly countries.
- The balance between effective regulation and stifling innovation and competition [*2 instances*].

## OPPORTUNITIES

Like the responses for barriers, the responses to the opportunities exercise are ranked by the number of times a specific opportunity was mentioned, or where several opportunities can be categorised together by the number of responses in that category. Responses were also filtered to keep only those that relate to security and safety of autonomous systems, rather than AI in a wider context.

### TECHNICAL

- No technical opportunities had any consensus.

### FINANCIAL

- No financial opportunities had any consensus.

### PEOPLE AND SKILLS

- Instilling a security culture in an organisation or at a national level (akin to openness around safety reporting) would be low cost and high impact [*8 instances*].
  - Key aspects of this include internal transparency and openness in organisations, clear accountability for failures, whistle-blower protection, and strong enforcement of policy.
- Creation of an industry-owned good practise guide, developed over time from experience [*4 instances*].
  - Suggestions that this could be driven by linking existing initiatives and organisations to form one coherent group on AI safety in the UK, and that the government could fund this.

- An opportunity exists in the willingness of stakeholders in the UK to contribute experience and solutions to other UK organisations [*3 instances*].
  - Specifically, the transfer of knowledge and skills between sectors, including from mature industries to emerging novel ones.
- That there exists a clear appetite for better and more transparent collaboration between industry and academia [*2 instances*].

## REGULATION AND POLITICS

- An opportunity to build on existing robust and effective standards rather than starting from scratch to give regulatory bodies an opportunity to keep up with the rate of technological progress [*3 instances*].
- The UK is in a strong position to influence or lead the creation of international standards [*2 instances*].
- The UK government has an opportunity to share information about big projects and their considerations for AI and security [*2 instances*].
- The public needs to be involved in considerations of AI and security, including on regulation, data, security, and privacy to ensure trust in autonomous systems [*2 instances*]

# WORKSHOP 4: PANEL DISCUSSION AND WRAP-UP

## OBJECTIVE

The final session of the workshop was an opportunity for all delegates to be in the same room and engage in discussion and feedback based on the event's three main workshops. The session was framed with a panel discussion session featuring **Dr Guy Burroughes**, **Dr Horia Alexandru Maior**, and **Dr Tarek Gaber**, and chaired by **Phill Mulvana**.

This session was designed to:

1. Elicit opportunities for collaboration and key challenges identified by the delegates over the course of the entire workshop and look for commonality.
2. Ask delegates for direction on next steps in the areas of AI assurance,
3. Offer delegates a chance to ask questions and offer insight to the entire delegation, including on topics which may not have specifically been touched upon previously.
4. Wrap-up the workshop and offer thanks to all those who attended and offered their insights and knowledge.



*Figure 11: Members of the discussion panel. From left to right, Dr Guy Burroughes, Dr Horia Alexandru Maior, and Dr Tarek Gaber.*

## KEY FINDINGS

The panel session enabled experts from a cross-section of the topics covered to respond to questions from the delegates, submitted digitally via Slido. Delegates could also submit their insights for opportunities for collaboration and key challenges via Slido, which created a word cloud of common terms.

## OPPORTUNITIES FOR COLLABORATION AND KEY CHALLENGES

The results of the question on possible opportunities for collaboration can be found in Figure 12 and Figure 13.

The opportunities for collaboration mostly focus around:

- A clear need for common standards for AI across different sectors.
- The role of organisations such as RAIUK and RAICo in facilitating collaboration, including between industry and academia.
- The multidisciplinary nature of AI, and the possibility of collaboration between sectors on solving challenges.

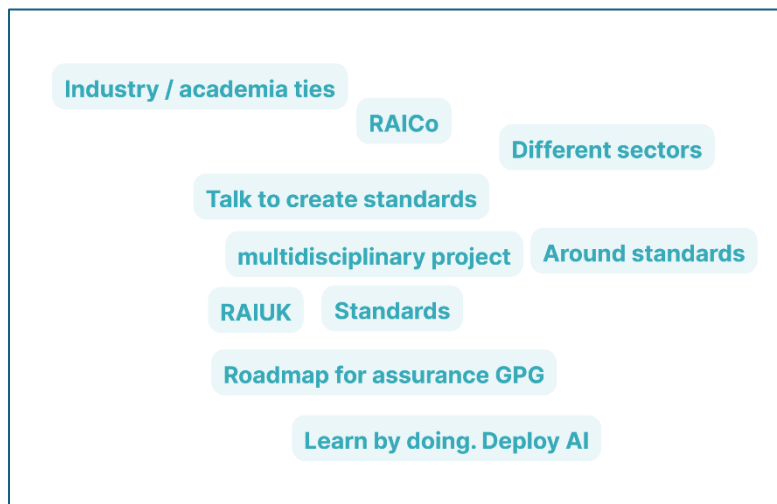


Figure 12: Word cloud for collaboration opportunities. 17 responses from 11 participants, with common ideas highlighted.

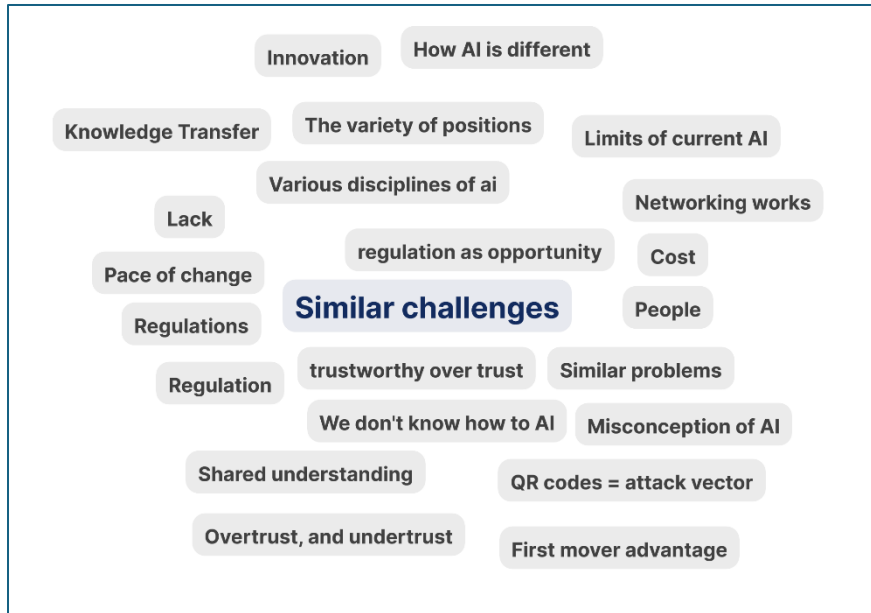


Figure 13: Word cloud for key insights developed over the workshop. 38 responses from 23 participants, with common ideas highlighted.

The key insights mostly focus around:

- The similarity of challenges and problems in the use of AI across sectors and domains. This was overwhelmingly the most common response.
- The importance and role of regulation in the use of AI.
- A lack of understanding of what AI is, how it can be used, and its limitations.

## NEXT STEPS

Delegates were asked what the next steps should be after the workshop, with regards to AI assurance, AI ethics and responsible innovation, and AI security. The responses were collected via Slido, and common themes are discussed below.

### ASSURANCE

- The creation of cross-domain industry-led guidance and good practice for assurance in AI, including a taxonomy of assurance techniques and technologies.
- A review of what lessons could be learned from non-AI technologies, which could be applied to assuring AI systems.
- More education for stakeholders as to the importance of AI assurance.



## ETHICS AND RESPONSIBLE INNOVATION

- Educate stakeholders on the importance of ethics and responsible innovation in the context of AI – particularly to waylay concerns about hindering innovation or progress.
- Create guidance and use cases to teach stakeholders about when ethics is a consideration for AI development or deployment, and guidance on what to do to implement RI.
- Include a wide spectrum of voices and stakeholders into conversations on ethics and RI for AI – what is ethical and responsible for AI development is not as clear as in other domains.

## SECURITY

- Find ways to improve education and awareness around the security risks posed by AI, both in terms of autonomous systems used within an organisation, but also AI used as a method of attack.
- Foster a security culture in organisations that instils the importance of security awareness, but also ensures people can raise security concerns with confidence they will be taken seriously.
- Like assurance, the creation of cross-domain industry-led guidance and good practice for security in AI, including a taxonomy of security techniques and technologies.