CfAA and RAICo Workshop on Safety of Robotics and AI in the Nuclear Domain

Workshop Report

Dr Richard Hawkins
Dr Calum Imrie
Dr Ioannis Stefanakos
Dr Sepeedeh Shahbeigi

Centre *for* Assuring Autonomy

Introduction

Overview of workshop

This workshop on the safety of robotics and AI in the nuclear domain arose out of the RAICo1-funded research project AID (Assuring the safe use of AI-enabled autonomy in nuclear Decommissioning). The AID project, lead by the Centre for Assuring Autonomy (CfAA) at the University of York investigated the how approaches to safety assurance for AI-enabled systems can be applied and adapted to nuclear decommissioning applications to support their adoption into nuclear facilities. This is a large, challenging and multifaceted problem with multiple stakeholders such as system and technology developers, operators and regulators. To support the validation of the technical outputs from the project, and to consider the way forward and ongoing challenges in this area it was therefore recognised that it would be important to bring together representatives from different stakeholder groups with different views and perspectives.

The workshop was split into two main sessions. The first session focused on knowledge-sharing through presentations from the project team at York as well as invited speakers. This provoked interesting questions and discussions, and the outputs from Mentimeter polling of the attendees is attached to the report. The second session involved interactive group discussions on topics that were prioritised by the workshop attendees. These sessions resulted in some very useful output identifying the key challenges as seen from the represented community, and the possible approaches for addressing these. The outputs from these group sessions is summarised in this report.

Workshop attendees

The workshop had attendees from both academia and industry. These included the University of York, University of Manchester (RAICo1), University of Warwick, Babcock, UKAEA, Sellafield, Cavendish, Office for Nuclear Regulation, and Sellafield. There was expressed interest from other groups, such as the National Nuclear Laboratory, but unfortunately were unable to attend on the day. They have though enquired about the outcome of the workshop, and how, as collective, we can continue these discussions to facilitate the safe integration of autonomous systems into the nuclear domain.

[1] RAICo, the Robotics and AI Collaboration is a collaboration between the UK Atomic Energy Authority (UKAEA), Nuclear Decommissioning Authority (NDA), Sellafield Ltd and the University of Manchester.

Breakout groups

The morning session had the attendees identifying and selecting themes that were to be discussed in the afternoon. There were three breakout groups, one for each selected theme, to consider the challenges associated with their assigned theme. Using their expertise and knowledge of working within the domain, the attendees shared insight of how to pursue these challenges and what would be required as a collective. Each group had different approaches to discussing and presenting their discourse to the other groups, which have been maintained in the following summary notes of each breakout group.

Constructing Assurance Arguments

It is paramount that prior to deployment the assurance of an AI-enabled systems is rigorously constructed. Below are key issues raised for the various aspects of assuring AI systems:

1. **Evaluating AI's behavioral safety**
    - How feasible are the available analytical approaches for techniques like machine learning?
    - How do you judge whether the data used in developing ML systems is good enough from a safety perspective?
    - What are appropriate metrics for AI safety that can be used in the safety case?
        - How will acceptable performance of the AI be judged?

2. **Assurance of AI-enabled system**
    - How much of the claim of safety is it acceptable and appropriate to prescribe to the AI?
    - How do you make sure that there is a sufficient correspondence between the training data and the data inputs that will occur in the operational environment?
    - How is a body of good practice established for AI safety assurance that can be used to help make judgments on acceptability?

3. **Creating safety cases**
    - How do we deal in the safety case with the unpredictability of AI failure modes?
    - The safety case cannot rely on evidence from trial and error alone, but what other viable means of verification evidence are available?
    - There could be (or could be perceived to be) a high cost associated with creating the required arguments and evidence for the safety case. How is this managed so that it doesn't become a barrier.

4. **Assessment of safety cases**

    - Uncertainty around the acceptance of safety cases for AI means there is a high risk to the project of adopting such an approach (and a high cost to failing to achieve approval).
    - There may be a lack of skills on the part of the assessors or regulators in terms of how to review and assess safety case arguments (this can also feed into the uncertainty around acceptance).
    - Will organisations have IP concerns over sharing of information as part of the safety case?

Architecture of ML Systems for Safety Performance

Designing ML systems for safety-critical applications presents unique architectural challenges. Below are key issues and recommended approaches for addressing them:

1. **Data Requirements**

    - *Challenge:* ML systems rely heavily on real-world data, which may be incomplete, biased, or hard to obtain.
    - *Approach:* Combine observational data with synthetic data generation techniques. Use synthetic data to explore edge cases, but validate models carefully to avoid extrapolation errors.

2. **Scenario Coverage**

    - *Challenge:* Capturing and modeling all possible operational scenarios is inherently difficult.
    - *Approach:* Employ adversarial neural networks and robustness analysis to uncover weaknesses. Adopt robust training methodologies to ensure reliable performance under varying conditions.

3. **System Requirements**

- *Challenge:* Vague or evolving requirements can lead to unpredictable system behavior.
- *Approach:* Establish a continuous requirements refinement process involving all stakeholders. Clearly define constraints and align model complexity accordingly.

4. **Standardization**

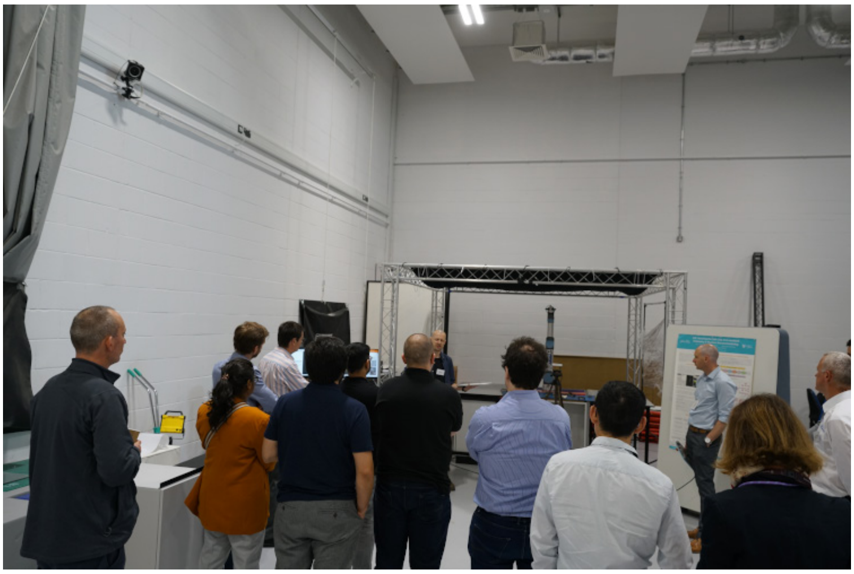
- *Challenge:* Lack of standardized architectural patterns and deployment practices.
- *Approach:* Develop a modular and transparent ML pipeline—from data logging to system deployment. Promote use of templates and guidelines, balancing openness with protection of commercial interests.

5. **Performance Metrics**

- *Challenge:* Evaluating safety-related ML performance with generic metrics may miss critical behavior.
- *Approach:* Define task-specific and safety-relevant performance metrics. These should reflect real-world success criteria and be validated with domain experts.

6. **Meta-Architecture**

- Consider a “meta-architecture” approach—an overarching framework that supports the integration, modularization, and systematic evaluation of different ML components within a safety-critical system.



AI Safety and Operational Practices

The breakout discussion for group 3 was on the topic of AI safety and operational practices. A range of technical, organisational, and human factor challenges influencing the reception and safe integration of AI systems in nuclear domain was explored. Potential solution was also discussed. The summary of the discussions are as follows:

1. **Demonstration of cost/benefit of deploying AI:**

    - A need for transparency with the operational staff with tangible evidence on the advantages and disadvantages of using AI systems was raised.
    - It was recommended that the teams be included in discussions and decisions around AI deployment to foster trust and ownership.

2. **Understanding the roles/tasks adequately to allow successful AI deployment:**

    - It was noted that roles and tasks must be clearly defined and understood before AI is deployed.
    - Without this clarity, operational inefficiencies or resistance may be encountered.

3. **Liability in the event of a failure:**

    - The need for accountability frameworks was identified to ensure responsibility is clearly assigned when AI-related failures occur.
    - Legal, organisational, and procedural aspects were considered critical.

4. **Understanding the acceptable level of failure and what is the definition of failure:**

    - It was acknowledged that not all failures are equally critical.
    - Definitions of failure and acceptable levels of risk or uncertainty should be established.

5. **Attempting to make AI mimic the human:**

    - The assumption that AI should behave like a human was challenged.
    - It was noted that human behaviour may not always be an appropriate benchmark for AI system performance, particularly in high-assurance or safety-critical applications.

6. **Education / Communication / Selling the benefits / Define the limits:**

    - The importance of structured engagement with operational teams was highlighted.

- It was recommended that training, communication of benefits and limitations, and involvement in decision-making processes be incorporated to support acceptance and safe integration.

Going forward

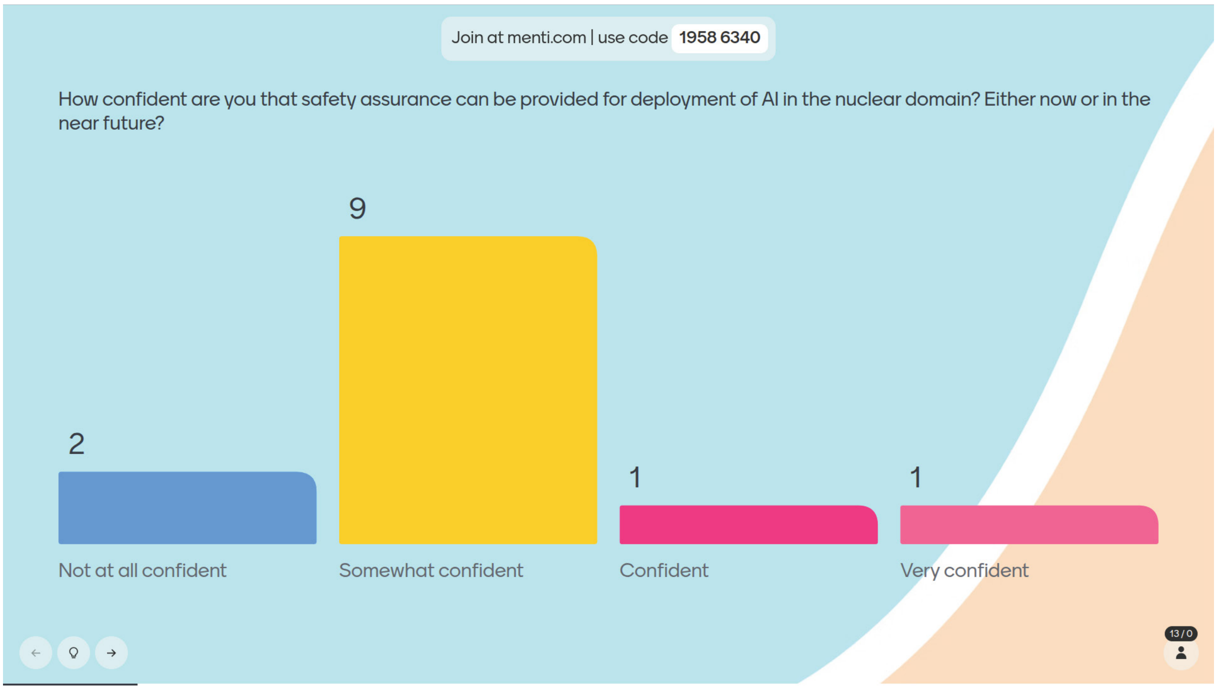
The workshop demonstrated the commitment the wide variety of stakeholders within the nuclear sector have for ensuring the safety of AI and autonomous systems when deployed for safety critical applications. The attendees were keen to engage with each other, and there was agreement that collective effort will greatly assist in enabling autonomy for nuclear operations. Follow up workshops like this will be key for bringing together stakeholders to continue constructive discussions, share new insights, and establishing end-to-end demonstrator projects.

We thank all the attendees for their active participation at the workshop. If you would like to know more about the workshop or how we can establish collaborations please do not hesitate to contact us.

Appendices

Results from Mentimeter questions

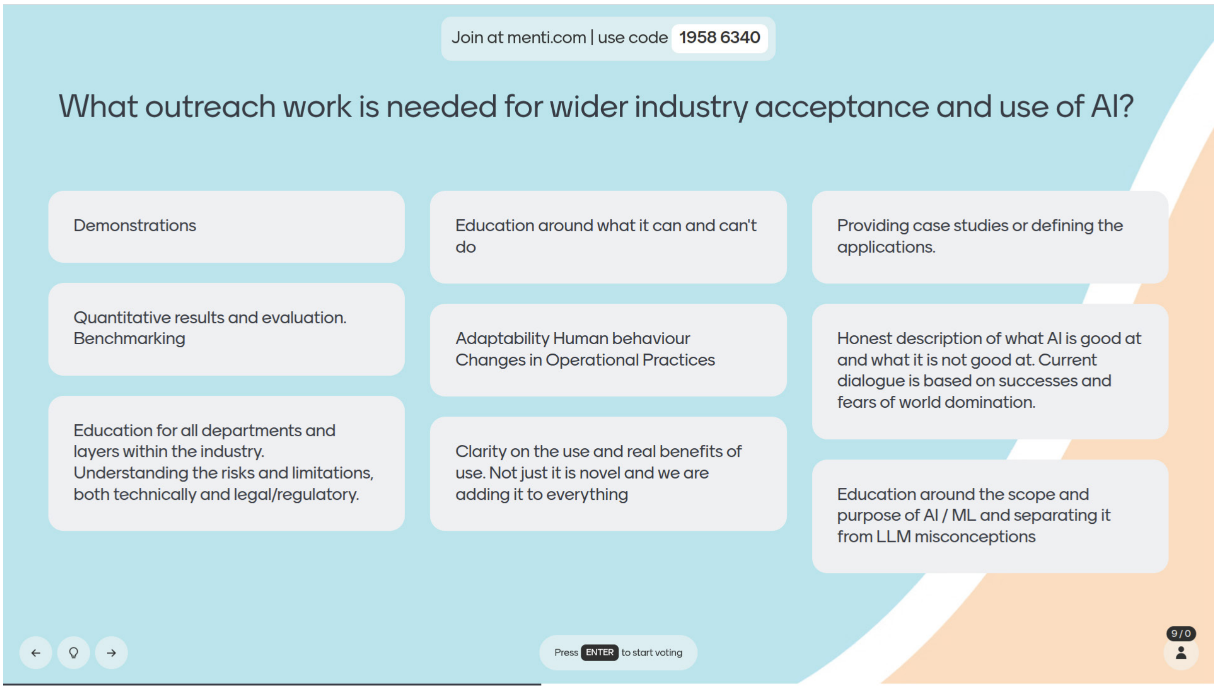
Q1. How confident are you that safety assurance can be provided for deployment of AI in the nuclear domain? Either now or in the near future?



Join at menti.com | use code 1958 6340

How confident are you that safety assurance can be provided for deployment of AI in the nuclear domain? Either now or in the near future?

Q2. What outreach work is needed for wider industry acceptance and use of AI?



Join at menti.com | use code 1958 6340

What outreach work is needed for wider industry acceptance and use of AI?

Demonstrations

Quantitative results and evaluation. Benchmarking

Education for all departments and layers within the industry. Understanding the risks and limitations, both technically and legal/regulatory.

Education around what it can and can't do

Adaptability Human behaviour Changes in Operational Practices

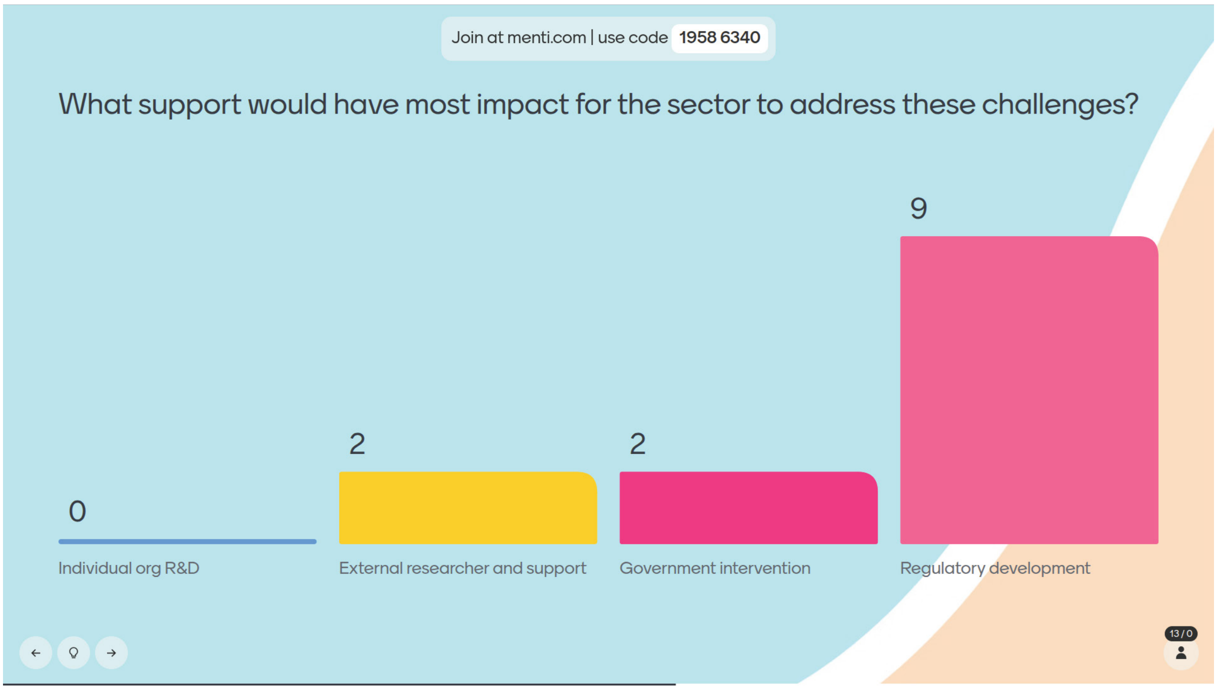
Clarity on the use and real benefits of use. Not just it is novel and we are adding it to everything

Providing case studies or defining the applications.

Honest description of what AI is good at and what it is not good at. Current dialogue is based on successes and fears of world domination.

Education around the scope and purpose of AI / ML and separating it from LLM misconceptions

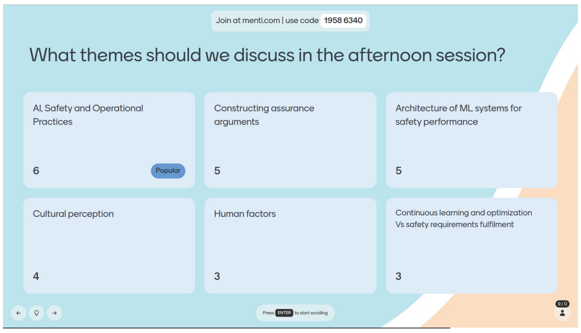
Q3. What support would have most impact for the sector to address these challenges?



Join at menti.com | use code 1958 6340

What support would have most impact for the sector to address these challenges?

Q4. What themes should we discuss in the afternoon session?



Join at menti.com | use code 1958 6340

What themes should we discuss in the afternoon session?

AI, Safety and Operational Practices
6 Popular

Constructing assurance arguments
5

Architecture of ML systems for safety performance
5

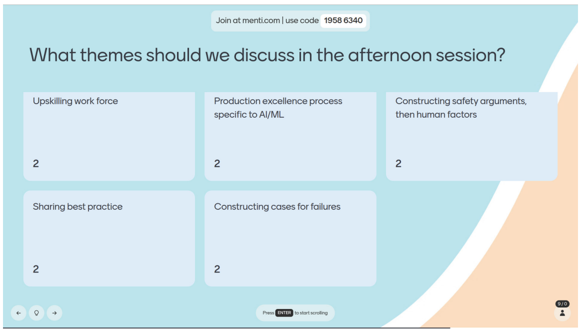
Cultural perception
4

Human factors
3

Continuous learning and optimization Vs safety requirements fulfilment
3

Press ENTER to start scrolling

9 / 0



Join at menti.com | use code 1958 6340

What themes should we discuss in the afternoon session?

Upskilling work force
2

Production excellence process specific to AI/ML
2

Constructing safety arguments, then human factors
2

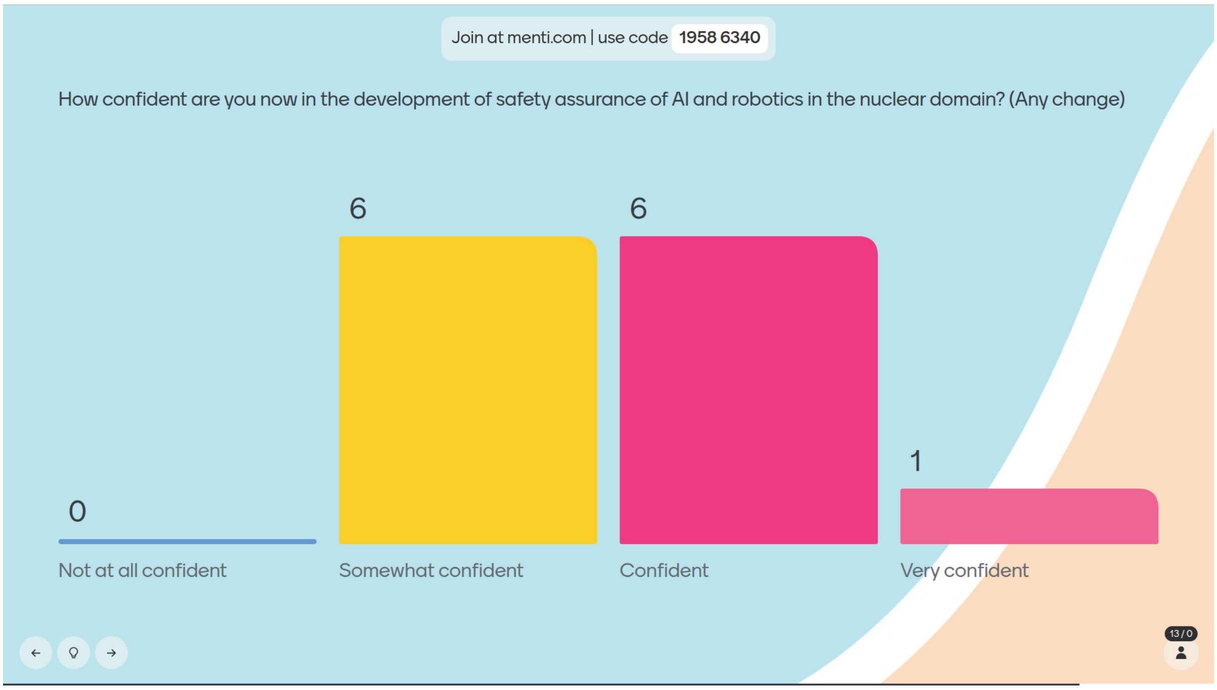
Sharing best practice
2

Constructing cases for failures
2

Press ENTER to start scrolling

9 / 0

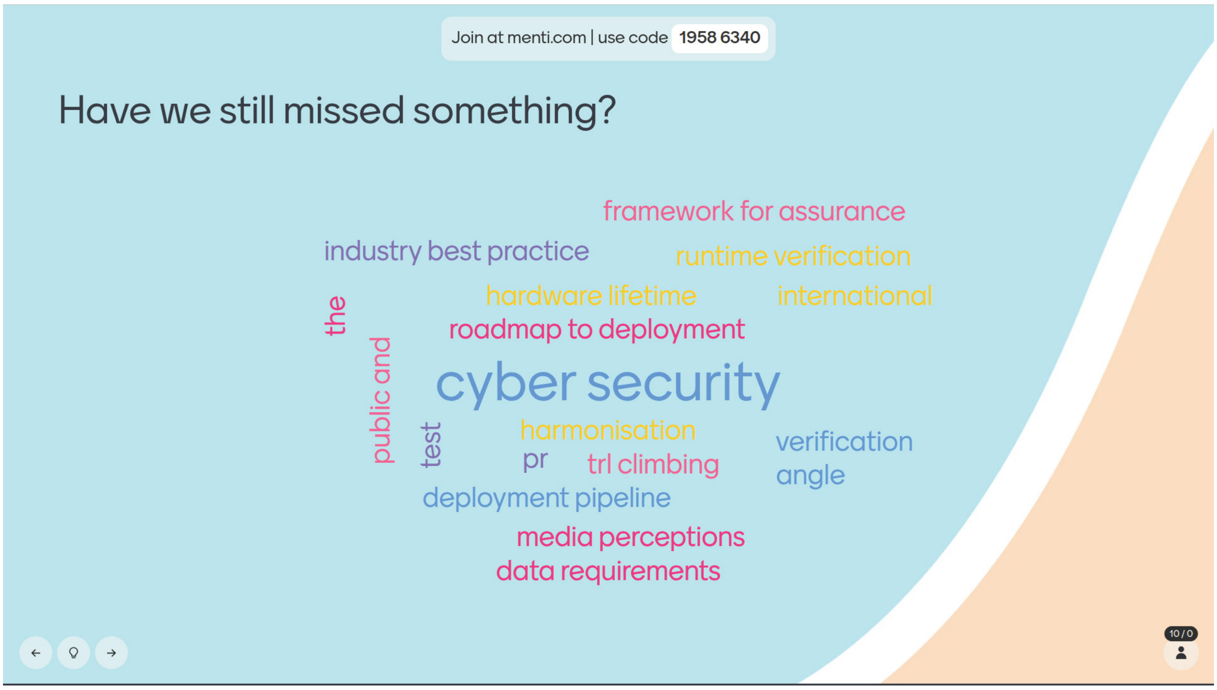
Q5. How confident are you now in the development of safety assurance of AI and robotics in the nuclear domain? (Any change)



Join at menti.com | use code 1958 6340

How confident are you now in the development of safety assurance of AI and robotics in the nuclear domain? (Any change)

Q6. What have we missed or not discussed today?



Join at menti.com | use code 1958 6340

Have we still missed something?

Contact

☎ +44 (0)1904 325345

✉ assuring-autonomy@york.ac.uk

linkedin.com/company/assuring-autonomy

Institute for Safe Autonomy,
University of York, York, YO10 5FT