AI and Operational Boundaries: A Basis for Analyzing Safety of Human-AI Teams

Andreas Bye
IFE, OECD NEA Halden HTO Project, Halden, Norway, Andreas.Bye@ife.no

Abstract: Artificial intelligence (AI) is coming, also in the nuclear industry. Three years ago AI and nuclear was not mentioned in the same sentence, while now even regulators embrace the technology. At the last Regulatory Information Conference by the U.S. NRC in Washington DC in March 2026, a human-AI session proposed that "AI and autonomous technologies redefine operational boundaries". What are these boundaries, and are the same questions and challenges relevant for all of them? Will AI be possible in the loop, and in which loop? This paper will outline some of the operational boundaries and discuss dimensions that need to be addressed in some of them. Human-AI teaming can be seen in relation to Human-automation, and this will also be discussed. Classic dimensions in human-AI teaming, such as explainability, trust and the role of humans will be discussed. Then a main question is how we can do safety analysis of such human-technology joint systems. What does this implicate for PRA? Can we use traditional HRA methods for the human part? Can we analyze the AI part for itself, or should the human-AI teaming be analyzed as one system? Which dimensions will be relevant for the various types of collaboration systems? This paper outlines the control loop and discusses a number of dimensions that need to be evaluated and a number of research questions to be addressed in the coming years.

1. INTRODUCTION

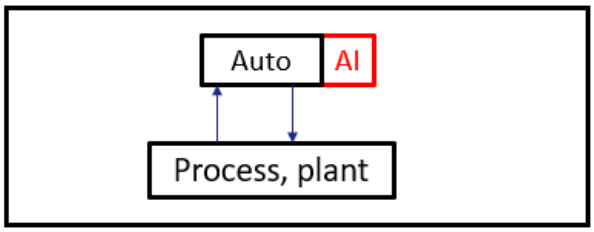
The nuclear sector is a conservative industry, and Artificial Intelligence (AI) has historically been viewed with caution. Only a few years ago, the answer was no to even mention AI in relation to nuclear applications. However, this has changed dramatically, and the question is now a more constructive one: What can AI be used for and how? AI is also at the forefront of a technology drive in society that should be handled with care. Are we going from being overly skeptical to being too optimistic without seeing the dangers and pitfalls of too fast deployment? Even regulators discuss the use of AI now, as can be seen in the last "sandboxing" exercises, e.g., the RegLab project at the Nuclear Energy Agency (NEA) [1]. At the Regulatory Information Conference (RIC) by the U.S. NRC in Washington DC in March, this was discussed in a session, posing the hypothesis *«AI and autonomous technologies redefine operational boundaries».* Before discussing the technologies and the opportunities for redefining operational boundaries, though, we should discuss what these boundaries are, and which challenges they have, in order to pose the right questions to evaluate the technology in each of them. This paper begins with a schematic overview of process control and discusses various dimensions that should be taken into account in each of these operational envelopes.

A central point when discussing types of technology and which types should be allowed in various operational boundaries, is the control loop. The term control loop is defined as containing the functions that directly influence the plant state in real time. Human operators are part of this control loop. Note that this control loop is specific for process control, which may be very different from other areas in which AI is used a lot these days, e.g., consumer product marketing. So the question "how to analyze safety of human-AI teams?" must begin with "when to analyze safety of human-AI systems". This paper outlines some factors that should be the basis for analyzing the safety of AI applications in nuclear process control.

2. THE CONTROL LOOP AND AI USE

Let us start with the inner loop, with "simple" control loops for example with classic control algorithms such as P- or PI- or PID-controllers controlling the level of a tank. This is normally called process automation or industrial automation. See Figure 1 for a schematic overview of this inner loop.

Figure 1: Basic process automation, schematic



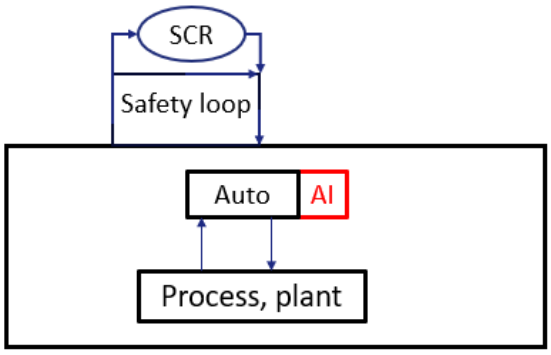
At the RIC conference in 2024, there was an example of a proposed AI application for process control, where a Machine Learning (ML) agent was implemented to control the level of a tank, in the chemical industry. It was tested since a classic PID-controller could not be used. Once it was trained, the ML agent worked perfectly [2]. What will be required for such an ML agent to be used in a nuclear setting? What first comes to mind, is **confinement**. If the controller can be confined with a limited input and output, then the algorithm can be tested for its intended use, and one may trust the use of it. A main question is whether there are guardrails for any consequences if the system or component fails. This can be assured in a better way if the system is confined.

In nuclear, the risk triplet is key. As stated by the NRC [3], the set of three questions defines the risk:

"1) What can go wrong?
2) How likely is it?
3) What are the consequences?"

This is the core of Probabilistic Risk Analysis (PRA). Used on the level of the ML-controller above, the two first questions can be resolved by confinement and testing. The third question if the controller doesn't work, is solved by back-up safety systems. This is the next step of operational boundaries in process control, see Figure 2.

Figure 2: Process automation with safety loop

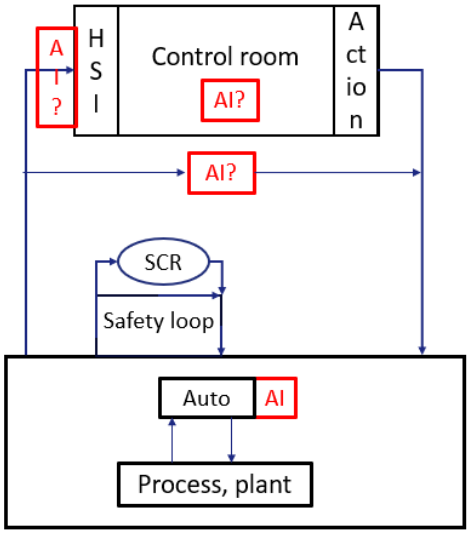


The safety loop takes care of emergency shutdown in case something goes wrong in the plant. Traditional dimensions for securing the functionality of these systems are diversity, analog backup, and formal methods. In the nuclear industry there has been a skepticism towards digital instrumentation & control (DI&C) for these systems as well. So this field does not seem ready for introducing AI based methods. However, there are thoughts about utilizing AI in architectures in such a way that the use of the AI is still safe. Can AI be used safely even if the AI itself is not safe? R. Winther discusses this in a new proposal [4]. In some way, this is similar to the use of AI or ML within the process automation. If

the system is confined in a known way, and there are guardrails against consequences, one might utilize AI for many purposes. The question is the extent to which safety claims are put on the AI system itself. If there are parallel systems or architectures, the safety claim may rest upon the traditional systems, not the AI.

The next step in the control loop, is the full loop with the human in the loop, in a nuclear power plant in a central control room, in addition to local controls in the field. This is sketched in Figure 3. Several possible AI applications are drawn here. The first is on the input side of the Human System Interface (HSI). One example of use is to use AI to filter alarms in case of disturbances. Alarm flooding has been a problem in process industries for many decades. Often, the operators use alarms in a pattern recognition manner in order to get an overview of the state of the plant. Thus, it might be an area in which even Large Language Models (LLMs) may be used. This might be an add-on advisory system for the operators in addition to their own analysis of the alarm situation. It is related to other kinds of systems in the control room. Many people talk about helping the operators diagnosing the plant state by the use of LLMs. Can LLMs support operators in understanding the plant state? This is an interesting question. What would be required of the LLM in that case? Will it have to be always right? Should it be used for exploration of hypothesized causes of failures? Should it be used for presenting final diagnoses? Will it be correct? Will the operators trust it, even when it fails? Won't the operators trust it when it is right? There are many questions here, many of them already posed in dealing with automation. Trust in automation has been a research issue for many years, exploring issues such as miscalibrated trust and automation transparency [5]. In short, this field of research includes all the human-automation research issues currently worked with in classic process control. In addition, it opens some new research questions.

Figure 3: The control loop of a nuclear plant



Human-AI teaming [6] is a new field, and core dimensions in this are: trust, transparency and explainability, situation awareness and sensemaking, and not least accountability and human control. "Human oversight" is a key concept even in new regulations such as the new AI Act from the European Commission [7]. What is this concept though, and when can it be said to be fulfilled, what are the requirements? Can there be human oversight without human insight?

AI systems in the control room can be of many types. They may be diagnosis systems, as discussed. They may also be procedure systems, semi-automated or more fully automated ones. In short one may use AI to support the operators in all their phases of operation, from detection through diagnosis and action planning to actions.

In Figure 3, there is an AI box bypassing the control room. This exemplifies AI used for a fully automated or autonomous system. It may be interrupted by operator actions though, following different levels of autonomy, e.g., as described by Parasuraman et al. [8]. Is a fully autonomous nuclear plant managed by AI thinkable in the near future? This again depends on the risk triplet. If the risk is acceptable it may be ok. A central point in this risk evaluation would be the worst-case consequences. This depends on factors such as siting and the details of the worst case scenarios. E.g., nuclear batteries in satellites, a remote plant under the earth's surface in a remote place, or even on the moon, might be acceptable. However, for normal industrial plants, the requirement for some kind of human oversight will probably still be required.

There are many dimensions of human-AI teaming systems mentioned already. Some of these means different things depending on from which angle you look, and different communities use the same words for different things. Trust is one case in point. Technology developers, including engineers making AI systems, often mean whether you can trust the outcome of an AI algorithm based on whether you can rely on the answer. I.e., the central point is that you can trust the system if the performance of the system is good (maybe with some degree of reliance). However, if you talk about trust from a user perspective, it means something different. A user can trust a system which gives wrong answers, and the user may distrust a system even if it gives correct answers all the time. This is called miscalibrated trust. The former meaning of trust used by engineers should rather be called trustworthiness. This is a capability of the system coming "from within" the AI system, meaning that it is a technological basis to trust the system or not.

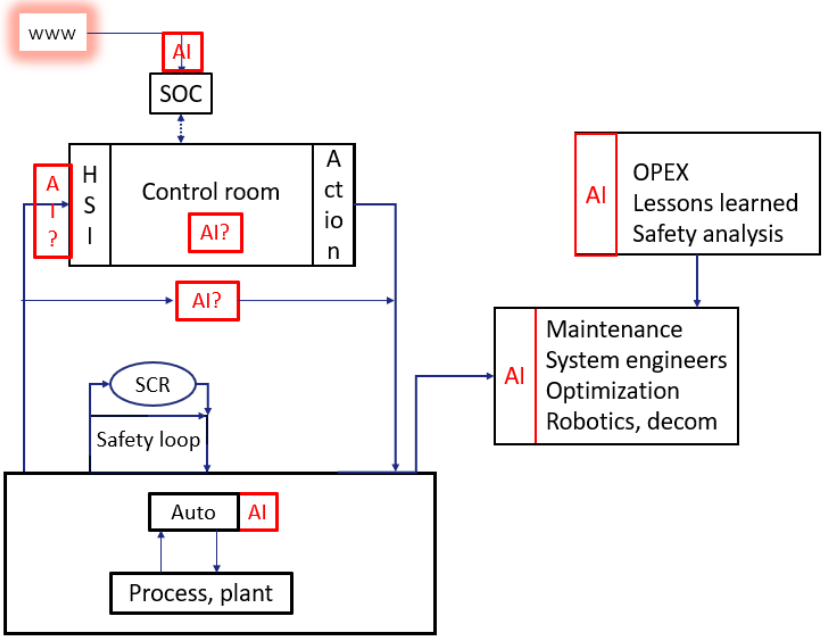
The same discussion can be made around concepts such as explainability and transparency. Skraaning and Jamieson described this related to automation as "seeing-into transparency" or seeing-through transparency" [8]. The former trying to explain the automation itself to the user, while the latter trying to explain the function or use of the automation. The former is something we don't expect from AI, it is considered a black-box related to its inner workings, and it is probably not possible to use a formal method to explain the inner functionality. However, the latter kind of transparency might be something to achieve even in human-AI systems, e.g., explaining the reasons and underlying judgement for the conclusions.

What about systems used in power plants external to the control loop? These are illustrated in Figure 4. In support for maintenance many utilities have already begun exploring AI. Condition-based maintenance has been a topic studied for years. In the Halden Project, this started in the 1990s with something at the time called "neuro-fuzzy systems". These were used for signal validation and instrument calibration [9]. Neuro-fuzzy systems would be called machine learning today, and many industries, including nuclear, is utilizing ML for condition-based monitoring these days.

The main difference between these systems and the systems in the control loop is the user. In the control loop an operator is using systems for online control of the plant, while in maintenance it is often system engineers who are the users. They are not responsible for the online surveillance. The offline nature of this use comes naturally with less regulations, and thus modern methods are more easily deployed. That said, there is a link between offline maintenance and online safety. Latent failures and weaknesses may be introduced in maintenance which have safety implications. However, the human oversight should be easier in this kind of off-line use since there is more time to analyse implications than in online use.

So ML is used for maintenance. What about LLMs? Currently there is a drive to utilize AI in very many offline applications. We see new applications for optimization and for new designs. This use of AI is often motivated by faster application and deployment time. This might be ok as long as the new solutions can be verified and validated in a way approved by regulations.

Figure 4: The control loop and off-line functions



The upper right box in Figure 4 illustrates yet another area of use, applying to operational experience reviews (OPEX) and learning from event reporting. LLMs have been used quite extensively for this in the last couple of years. LLMs enable searching through and categorizing lessons from a vast amount of written sources in a very short time. Some people questioned this, whether it was anything else than a strong search machine, and one good reply was: "it is like search on steroids". In addition, LLMs can group and organize findings in a way that few could have imagined only a few years ago. Back to the central question: Who is the user of this information? Normally, it will be off-line engineers or experts looking for ways in which to improve the plant. In this kind of use, the structured knowledge provided does not seem to have any regulatory concerns since it only increases the utility of the results.

However, there may be a new way of using this kind of system, coupling it to online advisory systems and using it to give online advice to operators in the control room. Or it could be used for online expert centres, in which experts can modify the process online, and make actions that directly impact the plant state. In that case, regulatory concerns apply, and it should be handled as one do with advisory systems inside the control loop, as discussed around Figure 3 above.

The upper left box in Figure 4 shows the Security Operations Centre (SOC), which is surveying and hindering cyber-attacks from the outside of the organisation. There is not a direct link between administrative systems and online plant systems in nuclear. However, there may be a weak link, hence there is a dotted line from the SOC to the control room. AI in all versions of the technology is used to defend against cyber-attacks. AI is also probably used to generate the attacks, so there the AI-war is going on already. This is included in the figure to illustrate that cyber security is a growing topic also in nuclear process control.

3. WHICH DIMENSIONS MUST BE EVALUATED

Summing up the use of AI in the different control loops, we can organize the discussion based on different viewpoints: First, AI is not AI. ML algorithms have different capabilities than LLMs, so regulatory constraints will be different for the different kinds of technology applied. ML algorithms can to a certain extent be validated to give the same output when given the same input, so in many ways they can be treated as deterministic software (still software though). LLMs have a completely different potential for hallucinating, meaning they can in principle not be trusted 100%. AI engineers may talk about correct answers 94% of the time, and this might be good response in some applications. However,

this does not hold for nuclear process control, which requires reliability in a complete different order of magnitude, talking about failure rates in the area of 10E-6. Whichever technology that will be used in nuclear, the responsibility of analysing the safety for the implied systems must be based on the same principles as they are now. So that does not change with AI. The risk triplet is still valid.

Another question is: When are we in the control loop? What kinds of actions have online impact? Are there hidden links, or can systems be confined and thus used for exactly that use that it was intended and designed for? There is probably no fixed answer to this question, but there are already a few links sketched above worth looking up for. One thing that has been long recognized is links between maintenance and latent failures introduced in the plant. In PRA and human reliability analysis (HRA), this is known as type A failures. Many analysts discuss whether these kinds of failures will be more dominant in the future. One reason for this may be the use of AI in design of plants and that assumptions that are put into the models do not hold and thus creating unknown latent failures that must be handled by online human oversight. Hopefully the operators will still have the knowledge and the means to handle them.

Related to the same discussion on hidden links is confinement. When is a system confined, and what are the criteria used to ensure it? Will there be other common cause failures that we don't think about now?

The dimensions on human-AI teaming mentioned above will be important to analyse: explainability, performance, risk, human oversight, trust and trustworthiness. When it comes to explainability, this is most often described based on the AI system itself. As argued above, I believe that that the more important dimension is the relation to the user of the system. Who is the user and what is the function of the output of the AI system? What will the user use the results for? This can be established in the framework of operational boundaries outlined in this paper. Establish the use, and then one can establish the requirements on the V&V of the system in use with the users. These requirements should not change based on the content of the technology. The same requirements on verifying the process of developing the system and validating the final product must be put on the total system independent of the technology inside.

4. THE ROLE OF HUMANS AND HOW TO SUCCEED

Humans have always had a strong belief in technology, and technology developments. The period we are in now is mixed, with some people warning about bad effects of AI. However, the technology drive in society seems stronger than ever, and the belief in what can be done with AI seems to have no limits. In the nuclear field, this comes at the same time as new reactor concepts are launched with more inherent, passive safety features and higher degree of automation. With AI on top of this, a relevant question may be whether there is a role for humans in the safe operation of future plants? My answer to that question is: Yes. Regarding automation, there will always be a role for the human in 24/7 safety related operations. The Boeing Max8 accidents is an example of what may happen if the belief in technology does not take humans into account. There will always be assumptions in the design phase that will not hold, and black swans will keep popping up. So humans must be in the loop. However, the role of humans will change. The first thing we see is that the role changes from manual operation to monitoring and review. This produces a set of new questions: Can we maintain the same assumption that humans will be capable of intervening after a long period of no actions, only monitoring? Will skills degrade? Will a new generation who has always done monitoring or reviewing, no actions or writing, maintain the capability of intervening or reviewing?

This might open for very pessimistic views on the future, if humans sink into passive roles and after a while lose even the ability to handle that passive role. However, are there other opportunities ahead? A wild idea might be to utilize the hallucinating weakness of LLMs to provide bad and good advice to operators in random sequences? Then humans will be forced to take a more active and interactive role with the system. Could we design systems to be really interactive, where hallucination is part of the strength, which will keep the operator on his/her toes but still provide constructive input? Maybe this

is a too wild idea, but the opportunities with modern technology are many, and could give us answers we have not imagined. Agentic AI is certainly a point in case here, to design agents that are truly human-centered and support the strengths of humans instead of making them passive spectators.

When new types of technology are introduced, with other capabilities than before, one important thing remains the same. The final requirements for safety and reliability do not change. Also, even if better tools are used for design, the process for how to evaluate the design and the safety of the design does not have to change. An important point that might be taken too easily is to establish the concept of operations early. In a technology driven world, the role of the human has often been considered in the final stage. However, if the decision is that the human shall be included in the control loop, e.g., to maintain so called "human oversight", this role must be designed into the system from day one. If this is done, one may make iterative testing and validation routines in the human factors engineering (HFE) process that incorporates the necessary requirements. In a collaboration with Idaho National Lab (INL), we have worked on the concept called Integrated Operations for Nuclear (ION) that aims to establish a model and concept of operations early in which the roles of all actors are considered.

5. SAFETY ANALYSIS OF HUMAN-AI SYSTEMS

How can we analyze the safety of new systems including AI? We can conclude that the safety requirements for nuclear power plants will still be very high, independent of the technology used. So will the assurance of the safety be based on different principles or dimensions than earlier? I think not. We still have to rely on PRA to answer the question "how safe is it?" And we still have to analyze both the process and the product when designing NPPs. Risk-informed and performance-based safety does not change in nature only because the underlying technology changes. In the details there are many interesting nuances though. When looking at HRA, the new technologies may include new failure modes. As discussed in [10] and [11], digital I&C and automation may create new failure modes that must be taken into account when doing HRA. However, will automation or support systems implemented with AI introduce new failure modes compared to automation? Maybe not, since it is the human role towards technology that is important, not the implementation of technology. If the HRA method is taking the human as a starting point, which most modern HRA methods do, they can still be used to analyze the human role in relation to the new technology.

6. CONCLUSION

AI has a lot of potential, and most actors in the nuclear industry try to harvest opportunities. This paper has outlined the control loop as a central framework and discussed the use of AI within and outside the control loop. There are different dimensions that are important inside the control loop vs outside. The nature of the human interaction with AI is also different in an online vs an offline setting.

The main conclusion is that AI cannot be let loose on safety systems or in the control loop without human oversight and control. The way to include AI is then to design good human-AI teaming solutions and test these before implementation. Make the concept of operations early, test it in simulated conditions, verify and validate. In short this is good Human Factors Engineering and risk-informed and performance-based processes.

The requirements for safety applications in NPPs won't change just because the technology is different. To be able to use AI in nuclear operations, especially within the control loop, there are a lot of research questions to address. With proper handling of these and making sure potential consequences are within acceptable limits, the possible opportunities are many.

On a last comment, the statement referred in the abstract was *"AI and autonomous technologies redefine operational boundaries".* In an attempt to write this paper very fast, I put the abstract into Copilot, and asked it to write out a paper. The main conclusion of this paper was exactly that AI redefines the operational boundaries. It justified this statement with a lot of capabilities of AI and sort of "beliefs of

the future". So, thinking about it once more, I do not agree with that. I believe that we must define the operational boundaries and operational envelopes first. Then we must come up with the different requirements for the various envelopes based on the risk. This will then result in different requirements to the technology, including the AI used, in the different operational envelopes. These requirements should build on many of the dimensions outlined in this paper, and others. So, disagreeing with my AI agent writing a proposal for my paper, I scrapped it totally and did not use any of it. Maybe next time I will agree with the AI agent, especially if I have been convinced by, I am sure, a lot of AI-generated talk in the time to come.

Acknowledgements

This work was supported by the OECD Nuclear Energy Agency (NEA) Halden Human-Technology-Organization (HTO) Project. Thanks to the members of the Programme Review Group of the Halden HTO Project for review and constructive comments.

References

[1] OECD/NEA. "*International Nuclear Sandboxing RegLab #1 – Final Report",* eno-040, Version 1 – Ennuvo Website Version" (2026).
[2] E. van der Bijl. *"AI Enabled Industrial Autonomous Operation in the Chemical Industry".* Presentation at the session W16 The Future of Nuclear: Adapting to AI-Enabled Autonomy, U.S. NRC's Regulatory Information Conference (RIC) (2024).
[3] https://www.nrc.gov/reading-rm/basic-ref/glossary/probabilistic-risk-analysis
[4] R. Winther. *"Safe AI vs Safe use of AI",* Proceedings of the 35th European Safety and Reliability& the 33rd Society for Risk Analysis Europe Conference. Edited by E. B. Abrahamsen, T. Aven, F. Bouder, R. Flage, M. Ylönen. ©2025ESREL SRA-E2025 Organizers. ResearchPublishing, Singapore. doi: 10.3850/978-981-94-3281-3_ESREL-SRA-E2025-P7886-cd
[5] G. Skraaning, G.A. Jamieson. "*Human Performance Benefits of The Automation Transparency Design Principle: Validation and Variation",* Human Factors, Vol. 63, Issue 3, pp. 379-401 (2021). https://doi.org/10.1177/0018720819887252
[6] National Academies of Sciences, Engineering, and Medicine. 2022. "*Human-AI Teaming: State-of-the-Art and Research Needs*". Washington, DC: The National Academies Press. https://doi.org/10.17226/26355.
[7] https://artificialintelligenceact.eu/
[8] R. Parasuraman, T. B. Sheridan and C. D. Wickens. "*A model for types and levels of human interaction with automation,"* IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans, vol. 30, no. 3, pp. 286-297, May 2000, doi: 10.1109/3468.844354.
[9] P.F. Fantoni, M.I. Hoffmann, R. Shankar, E.L. Davis. *"On-line monitoring of instrument channel performance in nuclear power plant using PEANO".* Progress in Nuclear Energy, Vol. 43, Issues 1–4, pp. 83-89 (2003).
[10] A. Bye, J.A. Julius, R. Boring. (2022). *"Challenges for Human Reliability Analysis in New Nuclear Power Plant Designs*". Proceedings of the 16th Probabilistic Safety Assessment and Management Conference, PSAM 16, June 26-July 1, 2022, Honolulu, Hawaii.
[11] A. Bye. (2024). *"Human Success: Old wine in new bottles, or a shift of mindset for HRA in an automated world?"* Proceedings of the 17th International Conference on Probabilistic Safety Assessment and Management & Asian Symposium on Risk Assessment and Management (PSAM17&ASRAM2024). 7-11 October, 2024. Sendai, Miyagi, Japan.