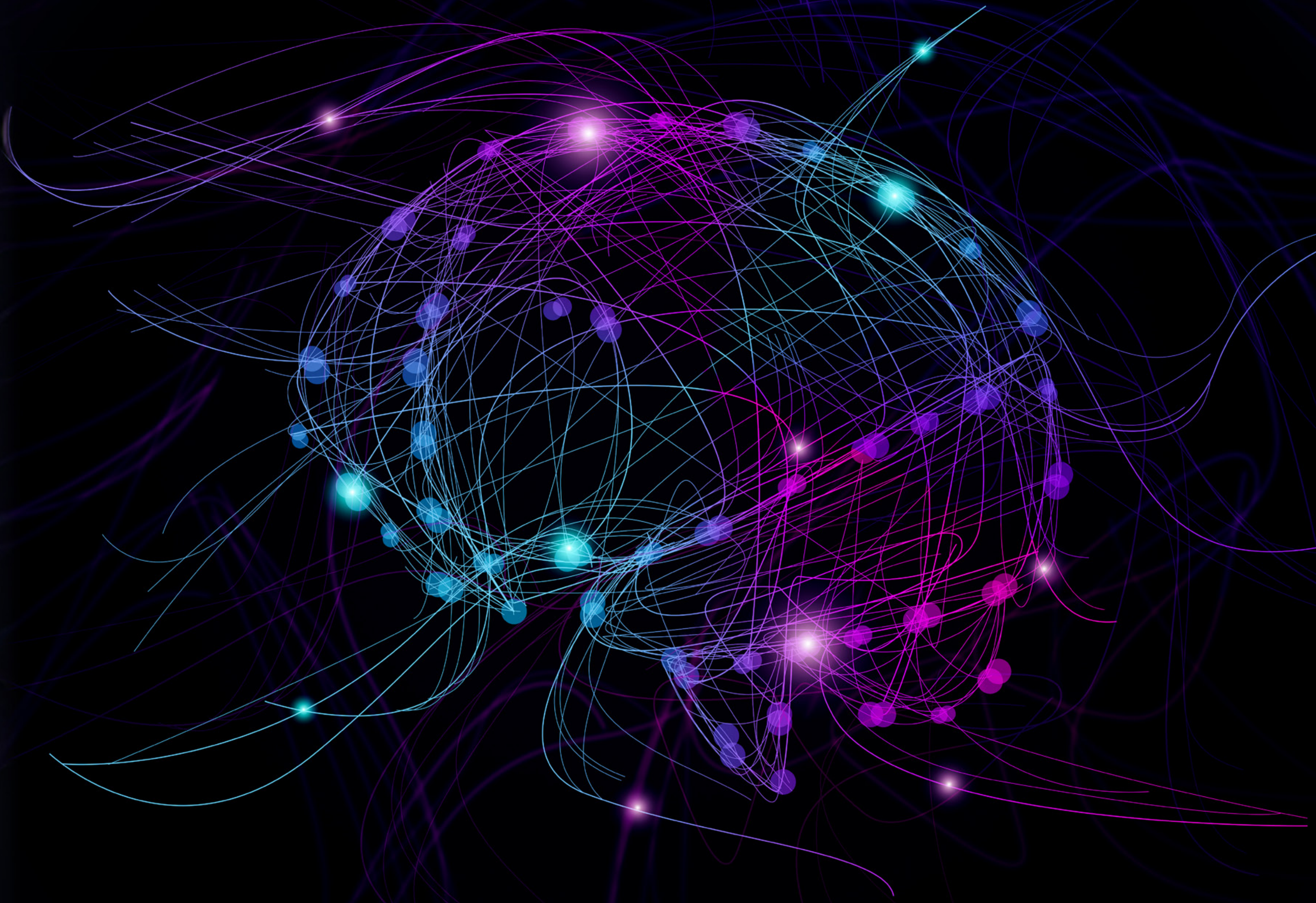
Emerging Technologies Workshop

Artificial Intelligence for Nuclear Verification

Workshop Report – Summary



IAEA

27–29 January 2025
Vienna, Austria

STR-406

Introduction

Artificial Intelligence is having a profound global impact. It has a vast range of applications across various sectors, and the field of nuclear safeguards is no exception. AI presents both opportunities and challenges for nuclear verification, and the IAEA Department of Safeguards is committed to leveraging this technology to enhance both the effectiveness and efficiency of its safeguards activities. In line with its mandate under the safeguards agreements to “take full account of technological developments”, the IAEA Department of Safeguards is actively exploring ways to leverage AI in its safeguards processes and activities and advance its AI-related projects.

Kory Sylvester
Director, Division of Concepts and Planning

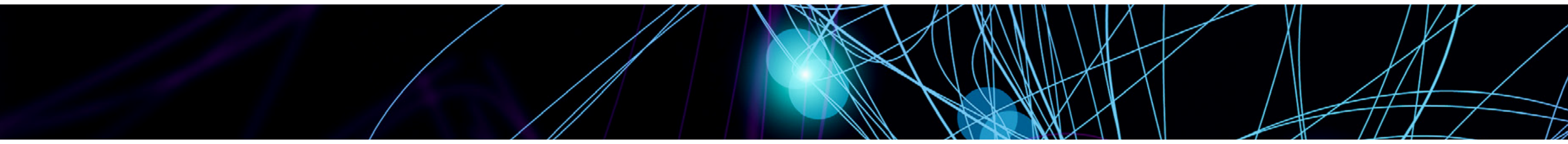
Stephane Baude
Director, Division of Information Management

The integration of AI into the verification activities are already bringing benefits, including enhanced analysis of data and safeguards relevant information and review of surveillance footage and video, resulting in more efficient use of resources. Further adoptions are expected to improve the Department’s monitoring and detection capabilities. AI advances may also change the proliferation landscape, requiring the Department to keep abreast of any developments that could undermine IAEA safeguards.

As part of its ongoing strategic foresight and planning activities, the IAEA Department of Safeguards held its third Emerging Technologies Workshop in Vienna, Austria, from 27-29 January 2025, bringing together 30 AI experts from diverse backgrounds, including academia, R&D institutions, government agencies, and the private sector to engage in in-depth discussions with IAEA staff. Member State Support Programmes, and the IAEA’s non-traditional partners attended as observers.

The Workshop highlighted the opportunities and challenges associated with AI adoption and emphasized the need for responsible use, careful consideration of key factors for safeguards activities, and collaboration with external experts. Finally, it identified actionable insights for the Department to consider as it moves forward with the integration of AI in its processes and activities.

The insights from the workshop will – inter alia – inform the Department’s suite of strategic planning documents, including the capabilities described in its Enhancing Capabilities for Nuclear Verification – Resource Mobilization Priorities required to meet its strategic objectives. They will also inform the internal policies and guidelines for AI for the Department of Safeguards.



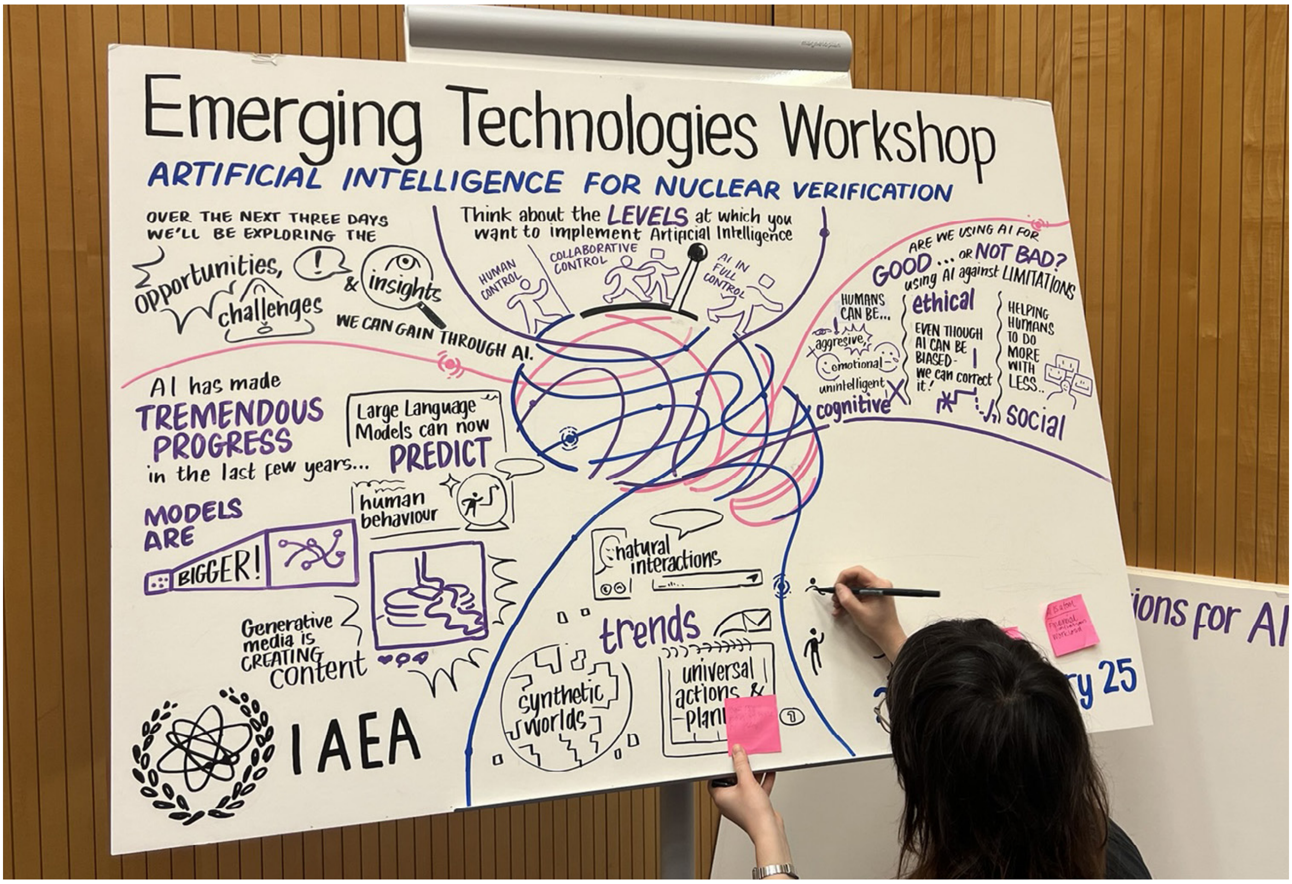
Executive Summary

The International Atomic Energy Agency (IAEA) Department of Safeguards (hereafter referred to as ‘Department’) held its third Emerging Technologies Workshop in Vienna, Austria from 27 to 29 January 2025. The workshop series aims to increase the Department’s awareness about and preparedness for addressing opportunities and challenges arising from emerging technologies by learning from external experts, and other organizations and sectors. The 2025 workshop brought together experts from various organizations, industries, and institutions to examine advancements in and applications of artificial intelligence (AI) and explore its potential for enhancing the efficiency and effectiveness of IAEA safeguards.

The workshop aimed to:

- equip Department staff with an understanding of recent advancements in AI and its future directions
- discover how other organizations and sectors leverage AI and mitigate potential risks
- create the knowledge base for informing the Department’s AI strategy and policies
- assist the Department to enhance and find solutions to its AI projects and applications

The event featured 30 external speakers and experts who shared insights on the current state of AI, highlighting its potential to revolutionize various sectors, and provided practical advice to the Department.


Emerging Technologies Workshop
ARTIFICIAL INTELLIGENCE FOR NUCLEAR VERIFICATION
OVER THE NEXT THREE DAYS WE'LL BE EXPLORING THE
opportunities, challenges & insights
WE CAN GAIN THROUGH AI.
Think about the LEVELS at which you want to implement Artificial Intelligence
HUMAN CONTROL
COLLABORATIVE CONTROL
AI IN FULL CONTROL
ARE WE USING AI FOR GOOD... OR NOT BAD?
using AI against LIMITATIONS
HUMANS CAN BE...
aggresive,
emotional
unintelligent
cognitive
ethical
EVEN THOUGH AI CAN BE BIASED - we can correct it!
HELPING HUMANS TO DO MORE WITH LESS..
social
AI has made TREMENDOUS PROGRESS in the last few years...
Large Language Models can now PREDICT
human behaviour
MODELS ARE BIGGER!
Generative media is CREATING content
natural interactions
trends
synthetic worlds
universal actions & plan
IAEA
ions for AI
y 25

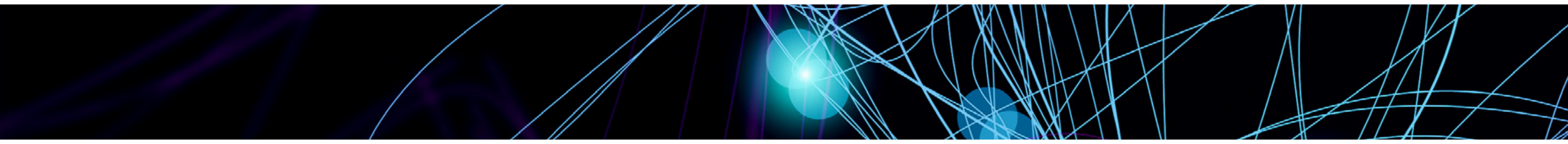
AI for Nuclear Verification – Key Insights and Ideas for Action

This report provides a comprehensive summary of the key trends and insights as well as actionable ideas that the workshop produced.

Part 1: Opportunities and Challenges for IAEA Safeguards

Key trends and insights

- AI has made significant progress in recent years in model sizes and capabilities: large language models (LLMs) can now perform multiple tasks, write code, produce plans, simulate human behaviours, and generate high quality video.
- Still emerging capabilities include the ability to build synthetic worlds and self-correct. As AI moves forward on the 'tool – consultant – collaborator – expert – autonomous agent' continuum and as it converges with other technologies, this will create new opportunities and risks.
- So far, organizations have deployed AI mostly for enhanced productivity and efficiency by automating routine tasks, freeing up human resources to focus on higher-value and more complex activities that require critical thinking and problem-solving skills.
- Organizations should experiment and try out where AI can be used effectively and start with small scale implementation, rather than wait for perfect solutions, and align AI with their business processes.
- AI can correct and overcome human limitations (e.g., memory), but it is does not replace human judgment and decision-making. AI outputs still need human validation.
- AI does not distinguish between correlation and causation. It does not reason but generates patterns from data. AI models require high-quality, relevant, and representative training data to function effectively. Errors, prejudices, and distortions in such data ('data bias') can result in unfair, inaccurate or misleading outputs that need to be effectively dealt with.
- Hallucinations – plausible sounding but completely wrong answers – in LLMs are not just occasional errors but currently an inevitable feature of generative AI, fundamentally tied to the probabilistic nature of these models.
- While AI's decision-making may not always be fully explainable (i.e., 'black box' problem), it is crucial that systems and their outputs remain rigorously validated to ensure trustworthiness.
- Auditing, ethical principles, documentation controls, governance frameworks, human-machine teaming, multi-layered guardrails, risk assessment, regulations, testing and evaluation requirements, and training help ensure responsible use of AI. Striking the right balance between speed and safety is key.
- The pace of change means that AI governance mechanisms can quickly become outdated. Ethical considerations are best incorporated into the organization's AI strategy and policies.



- Generative AI enables disinformation at an unprecedented scale. Deepfakes can take various forms, including images, videos, audio, and text. Fine-tuned AI models, fact-checking tools, regulations, and AI literacy can help combat this.
- AI may be used in nuclear material production, including in the sensitive stages of the nuclear fuel cycle (enrichment and reprocessing). Preventing misuse requires collaboration between policymakers and industries.
- AI is already offering effectiveness and efficiency gains in e.g., identifying and helping to prioritize safeguards relevant open-source information and reviewing surveillance footage. However, it cannot replace IAEA analysts and inspectors nor recommend safeguards conclusions.
- Safeguards specific considerations for AI include IT infrastructure and resource related limitations, information security requirements, quality assurance needs, as well as the need to maintain human control of safeguards processes and the resulting findings and conclusions.

Actionable ideas

- Align AI applications with business needs and organizational objectives and processes, and involve cross-functional teams in their development to ensure usefulness.
- Work on integrating information in IT systems and ensuring data quality in order to enable AI applications to fully leverage the organization's information assets.
- Ensure oversight of AI through governance frameworks, human controls and documentation.
- Guide AI development and use through policies and guidelines that align with the IAEA's legal mandates and ethical values.
- Adapt AI strategies and policies as AI technologies advance to address emerging risks and opportunities.
- Maintain human oversight and controls of AI systems and perform quality controls of AI outputs throughout their lifecycle.
- Start with small scale implementation, test and try AI solutions while assessing risks and quality, and put in place guardrails to ensure safe implementation of AI.
- In mitigating AI related risks, integrate AI considerations within the broader security frameworks.
- Maintain trust by ensuring AI applications align with the IAEA's mission and values and safeguards agreements, and through transparency and accountability.
- Promote AI literacy through training and education to enable staff to leverage AI's benefits and to ensure its responsible use.

IAEA
Emerging Technologies Workshop
...igence for Nuclear Verification
R. Blackman
G. G. Clavell
S. Baude

Part 2: Development of AI Tools for Safeguards Implementation

Key trends and insights

- Incorporating domain knowledge into LLMs can be achieved through the Retrieval Augmented Generation (RAG) approach, which provides references to source documents. In contrast, fine-tuning an LLM is resource-intensive and requires regular updates.
- "Knowing When You Don't Know" is a general problem in information retrieval, including RAG and LLMs; if you ask a question, there is always the risk of receiving an incorrect answer.
- Knowledge graphs help organize, analyse and validate large volumes of unstructured data, and enhance the capabilities of small language models that do not need significant computing power, thereby benefiting organizations such as the IAEA. They can provide new insights into textual data and enhance the factual accuracy of responses in RAG systems.
- Multimodal models introduce new opportunities for computer vision tasks (e.g., detection of unexpected movements on nuclear material in safeguarded facilities) and enable new applications such as image search and captioning. Zero-shot and few-shot learning with pre-trained multimodal models can provide useful results with minimal labelled data.

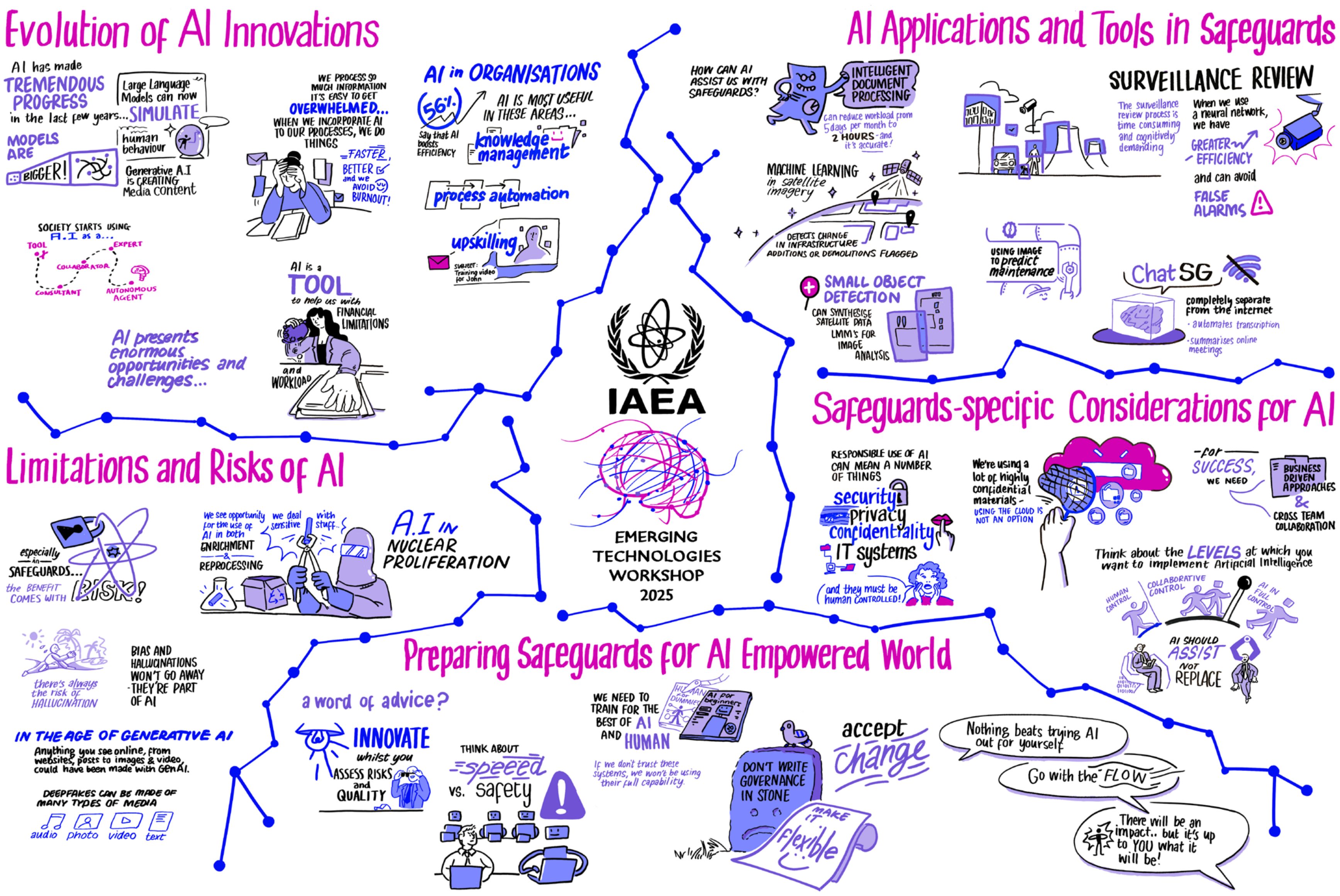
- Sensor fusion, which combines data from multiple sources, can provide deeper insights and more comprehensive analysis.
- AI agents introduce new opportunities to automate repetitive tasks but also introduce cybersecurity challenges, such as enabling large-scale social engineering attacks, and other security risks. Robust guardrails are essential to prevent unauthorized access, data manipulation, and unintended behaviour.
- Responsible AI considerations are essential for AI agents, requiring multi-layered safety measures to mitigate risks from non-deterministic behaviour, ensure transparency, and prevent unintended consequences.
- Understanding user needs and aligning AI with user expectations is crucial. Prioritizing solutions that address specific use cases while aligning with organizational needs remains a key factor for successful implementation.

  Smaller, targeted and applications and agents can often provide more reliable and efficient solutions than broad, general-purpose systems. Traditional machine learning and smaller models remain relevant for specific tasks, offering not only greater transparency but, in some cases, even superior performance.

Actionable ideas

- Explore the potential of multimodal models to detect unexpected movements in safeguards surveillance data.
- Investigate opportunities for AI to assist in predictive maintenance by detecting malfunctions in safeguards surveillance equipment.
- Consider traditional, smaller, and more transparent ML models for information retrieval and text classification tasks.
- Explore the potential of multimodal models to interact with visual content using natural language for tasks such as search or information extraction from large volumes of documents.
- Experiment with synthetic data generation to augment training data sets, while assessing its effectiveness in real-world scenarios.
- Develop internal benchmarks for satellite imagery analysis to evaluate model performance under safeguards-specific conditions and ensure reliable deployment in real-world scenarios.
- Be aware that hallucinations are a persistent property of LLMs and explore strategies for detecting and mitigating misleading outputs.
- Assess the potential of knowledge graphs to enhance domain representation and reduce factual errors in Generative AI applications.
- Adopt a problem-driven approach using simple, targeted models to solve well-defined tasks efficiently.

Workshop Summary Illustration



Evolution of AI Innovations
AI has made TREMENDOUS PROGRESS in the last few years...
MODELS ARE BIGGER!
Large Language Models can now SIMULATE human behaviour
Generative A.I is CREATING Media content
SOCIETY STARTS USING A.I as a...
TOOL
COLLABORATOR
CONSULTANT
EXPERT
AUTONOMOUS AGENT
AI presents enormous opportunities and challenges...
WE PROCESS SO MUCH INFORMATION IT'S EASY TO GET OVERWHELMED... WHEN WE INCORPORATE AI TO OUR PROCESSES, WE DO THINGS
=FASTER, BETTER and we AVOID BURNOUT!
AI is a TOOL to help us with FINANCIAL LIMITATIONS and WORKLOAD
AI in ORGANISATIONS
56% Say that AI boosts EFFICIENCY
AI IS MOST USEFUL IN THESE AREAS...
knowledge management
process automation
upskilling
SUBJECT: Training video for John
IAEA
EMERGING TECHNOLOGIES WORKSHOP 2025
AI Applications and Tools in Safeguards
HOW CAN AI ASSIST US WITH SAFEGUARDS?
INTELLIGENT DOCUMENT PROCESSING
can reduce workload from 5 days per month to 2 HOURS - and it's accurate!
MACHINE LEARNING in satellite imagery
DETECTS CHANGE IN INFRASTRUCTURE ADDITIONS OR DEMOLITIONS FLAGGED
SMALL OBJECT DETECTION
CAN SYNTHESISE SATELLITE DATA
LMM'S FOR IMAGE ANALYSIS
SURVEILLANCE REVIEW
The surveillance review process is time consuming and cognitively demanding
When we use a neural network, we have GREATER EFFICIENCY and can avoid FALSE ALARMS
USING IMAGE TO predict maintenance
Chat SG
completely separate from the internet
· automates transcription
· summarises online meetings
Safeguards-specific Considerations for AI
RESPONSIBLE USE OF AI CAN MEAN A NUMBER OF THINGS
security
privacy
confidentiality
IT systems
(and they must be human CONTROLLED!
We're using a lot of highly confidential materials - USING THE CLOUD IS NOT AN OPTION
-for SUCCESS, WE NEED
BUSINESS DRIVEN APPROACHES
&
CROSS TEAM COLLABORATION
Think about the LEVELS at which you want to implement Artificial Intelligence
HUMAN CONTROL
COLLABORATIVE CONTROL
AI IN FULL CONTROL
AI SHOULD ASSIST NOT REPLACE
Limitations and Risks of AI
especially in SAFEGUARDS... the BENEFIT COMES WITH RISK!
We see opportunity for the use of AI in both ENRICHMENT & REPROCESSING
we deal with sensitive stuff..
A.I IN NUCLEAR PROLIFERATION
there's always the risk of HALLUCINATION
BIAS AND HALLUCINATIONS WON'T GO AWAY - THEY'RE PART OF AI
IN THE AGE OF GENERATIVE AI
Anything you see online, from websites, posts to images & video, could have been made with GenAI.
DEEPFAKES CAN BE MADE OF MANY TYPES OF MEDIA
audio
photo
video
text
Preparing Safeguards for AI Empowered World
a word of advice?
INNOVATE whilst you ASSESS RISKS and QUALITY
THINK ABOUT speed vs. safety
WE NEED TO TRAIN FOR THE BEST OF AI AND HUMAN
HUMAN for DUMMIES
AI for beginners
If we don't trust these systems, we won't be using their full capability.
DON'T WRITE GOVERNANCE IN STONE
MAKE IT flexible
accept change
Nothing beats trying AI out for yourself
Go with the FLOW
There will be an impact.. but it's up to YOU what it will be!

Acknowledgements, Sponsorship and Support

The Workshop would not have been possible without the hard work of the following staff:

Strategic Planning Team: Jenni Rissanen (Team Lead, Scientific Secretary of ETW), Ilenia Bruseghello, Gleb Lebedev, Damian Meckbach, Louise Phillipson, and Loipa Sanchez Lorenzo.

Cross-divisional ETW Planning Team: Paul Schneeweiss (Co-Scientific Secretary of ETW), Stephen Bertl, Asim Hadzimujagic, Kristine Madden, Matthew Martinez, Stefano Passerini, Alex Pollack, Tatjana Stojadinovic, and Maikael Thomas.

Additionally, we thank the scene setters and session moderators including Rina Ahmed, Natasha Bajema, Stephane Baude, Sofya Hambaryan, Kelly Jenkins, Luisa Milani, Scott Partee, Kory Sylvester and Alexis Trahan. We also appreciate the support of the Department's communications team, IAEA graphic designers, and conference services.

The Emerging Technologies Workshop would not have been possible without the generous financial support of our Member State Support Programmes: Australia, France, Germany, Switzerland and the United Kingdom. They sponsored the participation of several external experts, provided various kinds of administrative support, as well as funded the scribed images that capture the workshop insights and ideas. Also, we would like to thank all other MSSPs for their in-kind contributions that helped to make this ETW a success, including that of Germany and Russia that provided human resources support in the form of CFEs and JPOs.

Disclaimer

This report was edited by the IAEA Department of Safeguards. The Proceedings have been edited by the editorial staff of the IAEA to the extent considered necessary for the reader's assistance. The views expressed remain, however, the responsibility of the named authors or participants. In addition, the views are not necessarily those of the governments of the nominating Member States or of the nominating organizations.

Although great care has been taken to maintain the accuracy of information contained in this publication, neither the IAEA nor its Member States assume any responsibility for consequences which may arise from its use.

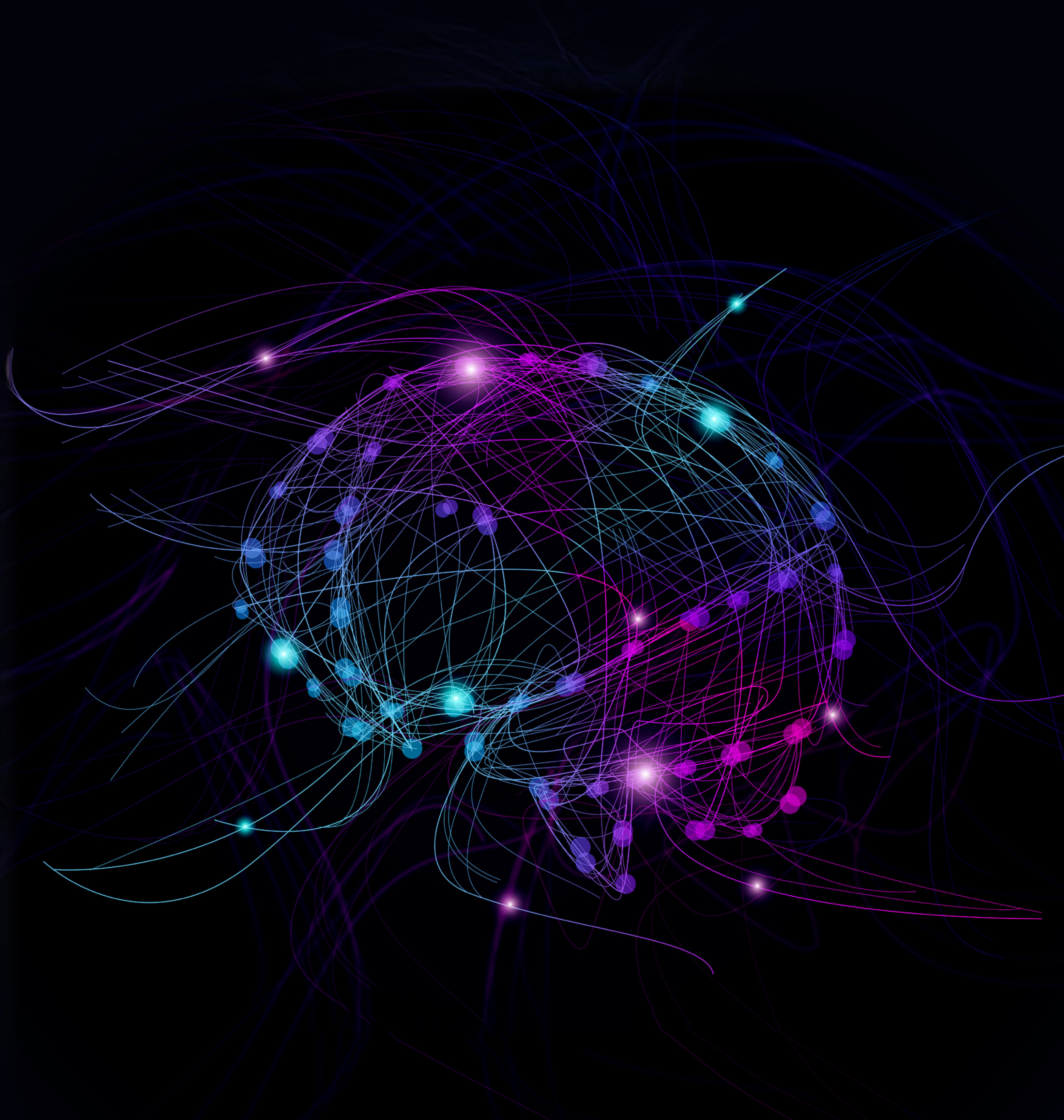
The use of particular designations of countries or territories does not imply any judgement by the publisher, the IAEA, as to the legal status of such countries or territories, of their authorities and institutions or of the delimitation of their boundaries.

The mention of names of specific companies or products (whether or not indicated as registered) does not imply any intention to infringe proprietary rights, nor should it be construed as an endorsement or recommendation on the part of the IAEA.

The authors are responsible for having obtained the necessary permission for the IAEA to reproduce, translate or use material from sources already protected by copyrights.

Material prepared by authors who are in contractual relation with governments is copyrighted by the IAEA, as publisher, only to the extent permitted by the appropriate national regulations.

IAEA



25-XXXXXE