ESTABLISHING ETHICAL GUIDELINES FOR APPLYING ARTIFICIAL INTELLIGENCE TO IAEA SAFEGUARDS

C.L. Murphy
Y-12 National Security Complex
Oak Ridge, TN. United States of America
Email: chantell.murphy@pxy12.doe.gov

J.L. Barr
Pacific Northwest National Laboratory
Seattle, WA. United States of America

Abstract

Drawing upon globally recognized efforts by a variety of institutions, the paper develops a practical and defensible framework for consideration and use by the International Atomic Energy Agency (IAEA) to support the adoption of credible ethical standards for autonomous and intelligent systems (AIS). The analysis is based on a future hypothetical AIS that identifies signatures of misuse, diversion, and other research and development (R&D). The hypothetical system serves as an exemplar for subject matter expert (SME) workshop participants to begin to answer ethical AIS questions around how these systems should be governed and maintained, what the internal and external considerations associated with artificial intelligence (AI) tools are, and where the boundaries between ethical and non-ethical uses are. The ethical AIS workshop explored several important topics:

(a) Recommendations that can be currently implemented by the IAEA,
(b) Proposed defensible criteria and metrics aligned with an ethical AI for safeguards implementation,
(c) Discussion of future needs and capabilities, and
(d) Derive appropriate next steps and actions to support filling current technological, operational, and cultural gaps and realizing an effective ethical AI for safeguards framework.

The Establishing Ethical Guidelines for Applying AI to IAEA Safeguards project creates awareness in the community of the ethical AIS approach.

1. INTRODUCTION

The Establishing Ethical Guidelines for Applying Artificial Intelligence (AI) to International Atomic Energy Agency (IAEA) Safeguards project initiated a study in 2020 to develop and apply a practical and defensible ethical AI evaluation framework to IAEA safeguards AI technology [1]. Employing such a framework will help the IAEA ensure autonomous and intelligent system (AIS) uses are accepted, or at least understood, by Member States.

One of the more robust frameworks considered, developed an ethical certification process with criteria for transparency, accountability, privacy, and algorithmic bias in AIS [2]. The certification process employs an efficient expert- and knowledge-driven method for quickly iterating the elements and variables of proposed AIS solutions that tend to affect a system's overall ethical qualities. An expert-driven approach may be attractive to the IAEA because of the commitment to being non-discriminatory and using consensus decision-making. Demonstrating that an AIS is ethically certified may reduce or completely remove hesitation by stakeholders about using AIS for safeguards applications.

The project team developed a comprehensive safeguards use case that uses AIS to find indications of diversion and misuse, including research and development (R&D) activities. The use case is the primary mechanism to elicit ethical requirements and metrics from subject matter experts (SMEs) in this project. The safeguards use case supports the final goal of the project, which is to deliver ethical verification criteria for AIS safeguards technologies that will promote ethical integrity in the full life cycle and regulatory processes within this emerging domain.

Because this is an evolving field, little is known or understood about the key dimensions pertinent to providing ethical verification of AIS safeguards technology. However, ethical guidelines are vital because it will

be extremely challenging to reconcile adopting AIS capabilities to improve the effectiveness and efficiency of safeguards with maintaining stakeholder trust by affirming that the algorithms are rigorous, transparent, and non-discriminatory. Given increasing global attention to ethical uses of AIS and the IAEA Department of Safeguards' interest in AIS, this study provides both timely and constructive information to support the IAEA's AI strategy development.

2. BACKGROUND

As identified in both the Safeguards R&D Plan [3] and "Development and Implementation Support Programme for Nuclear Verification, 2020–2021" [4], the IAEA is currently "identifying, evaluating, and testing" AI-based or automation-based capabilities to support open-source searches, network analysis, trade analysis, surveillance review, and continuous environmental scanning of the Department of Safeguards' operating environment. The automation and efficiency gain brought in by the AI-based or automation-based capabilities may also bring unintended and unethical consequences. Effective ethical AIS frameworks can help stakeholders identify critical needs in the areas of transparency, accountability, algorithmic bias, and privacy, develop strategies to address those needs, and ensure the appropriate mitigation efforts remain effective over time. The right ethical AIS framework will help stakeholders build trust in the AIS.

2.1. Current initiatives in AI and ethics

Safeguards literature does not address ethical uses of AIS, however, studies in other domains provide insight. It is common practice for private industries, governments, and organizations to create guiding principles documents or statements about the ethical use of AIS; it serves as a messaging tool to the world that they are committed to "ethical AI" (fair, unbiased, and transparent algorithms). According to a study published in *Nature*, only 9% of reports containing ethical principles for AIS come from international organizations, while the majority come from private companies and governmental agencies, followed by academic institutions [5]. However, few of these statements include actionable evaluation criteria and validation strategies.

Many U.S. corporations have adopted different approaches to address ethical AI, such as creating oversight boards and committees, issuing guidelines, and creating tools to operationalize ethical AI principles. Table 1 provides some examples of U.S. corporations that maintain organizational ethical AI guiding documents [5].

TABLE 1. CORPORATIONS AND THE TITLES OF THEIR ETHICAL AI GUIDELINES

Corporation	Title
Google	Our Principles
IBM	Everyday Ethics for Artificial Intelligence. A Practical Guide for Designers and Developers, IBM's Principles for Trust and Transparency
Intel Corporation	Artificial Intelligence. The Public Policy Opportunity, Intel's AI Privacy Policy White Paper. Protecting Individuals' Privacy and Data in the Artificial Intelligence World
Microsoft	AI – Our Approach, Responsible Bots: 10 Guidelines for Developers of Conversational AI
Unity Technologies	Introducing Unity's Guiding Principles for Ethical AI – Unity Blog

In addition to internal documents, several organizations have joined to develop industry best practices, like the efforts of the Partnership for AI to Benefit People and Society [6,7], and the World Economic Forum's *Empowering AI Toolkit* to help business leaders in their decision-making around AI [8]. In the U.S., the Department of Defense (DOD) and the U.S. Intelligence Community have developed and implemented ethical principles [9-11]. Many academic institutions have departments and centers to advance and promote AI research such as Stanford's Institute for Human-Centered Artificial Intelligence [12], MIT Media Lab [13], and the AI Now Institute at New York [14].

Internationally, the European Commission established the High-Level Expert Group on Artificial Intelligence (AI HLEG) in 2018 to develop *Ethics Guidelines for Trustworthy Artificial Intelligence (AI)* [15],

and in 2021 the United Nations published *A Framework for Ethical AI at the United Nations* [16]. The Institute of Electrical and Electronics Engineers Standards Association (IEEE SA) has several initiatives through its Global Initiative on Ethics of Autonomous and Intelligent Systems. Through its expert-driven approach, IEEE created the *Ethically Aligned Design (EAD): A Vision for Prioritizing Human Wellbeing with Autonomous and Intelligent Systems* document and established over twelve standards working groups inspired by EAD. The IEEE P7000 standards series addresses issues at the intersection of technological and ethical considerations identified by EAD by putting principles into practice for AIS [17]. IEEE SA also created the first certification process in 2018, the Ethics Certification Program for Autonomous and Intelligent Systems (ECPAIS), now called IEEE CertifAIEd™, which allows organizations to provide evidence for their due diligence to ensure their AIS products and services are ethically aligned [18]. In 2020, an IEEE SA team developed an ECPAIS use case for contact tracing apps (CTA) and contact tracing technology (CTT) to address how to verify greater ethical transparency, accountability, and privacy demonstrations in CTA/CTT [19]. In 2021, Vienna, Austria became the first city in the world to earn the IEEE CertifAIEd™ AI Ethics (AIE) Certification Mark to advance the city's Digital Humanism strategy [20].

Borrowing elements from different ethical AI initiatives like the Empowering AI Toolkit, the Ethics Guidelines for Trustworthy AI, and IEEE CertifAIEd™, the Establishing Ethical Guidelines for Applying AI to IAEA Safeguards project team developed an ethical AI safeguards use case. The safeguards AIS analyzes data from satellite imagery, literature and publishing sources, public records sources, social media data, and IAEA internal documents to find indications of diversion and misuse, including R&D. The use case is meant to be far-reaching to elicit SME discussion on transparency, bias, accountability, and privacy of such systems without being side-tracked with current policies, procedures, and availability of technology.

3. USE CASE AND ETHICAL AI FRAMEWORK

The safeguards use case is the primary mechanism to elicit ethical AIS requirements and metrics from SMEs. The use case design supports safeguards mission needs and improves current operations; it is forward-looking and broad enough to evoke discussion and requirements over the ethical AI landscape including bias, transparency, accountability, and privacy.

3.1. Ethical AI framework

The tailorable quality of the IEEE CertifAIEd™ criteria for certification in ethical transparency, accountability, bias, and privacy make the ontological specification reports attractive references to use for this study [21-24]. The ontological specification reports provide methods to assess and benchmark AI systems and organizations in their ethical performance regarding the key ethical principles of transparency, accountability, bias, and privacy. A set of goals and factors exist for each ethical principal and the degree to which those goals and factors are met influence the ethical performance of the AIS and organization. For example, some factors that enable ethical transparency, called drivers, are clarity and consistency of AIS operations and awareness of AIS interaction; and some factors that hinder ethical transparency, called inhibitors, are behavioral obfuscation and protection of trade secrets. The criteria documents also define ethical foundational requirements that describe actions, processes or structures that need to be in place to meet each goal, as well as evidence and metrics used to verify them. The goals, requirements, and metrics provide a pre-populated guide to create a tailored suite of recommendations for an ethical AI framework for safeguards AI technology. The process uses collective SME exploration of the AIS to generate the underlying ethical criteria or metrics that either impede or encourage the attainment of an ethically developed, operated, and maintained AIS.

3.2. Hypothetical safeguards AI use case

The hypothetical AI system developed in 2021 supports IAEA analysts by identifying signatures of misuse, diversion, and other undeclared nuclear fuel cycle R&D in 2031 [1]. The AIS described in this use case comprises sub-systems, or engines, used as part of the complete system or on their own. The system analyzes satellite video to identify the goods, services, and intellectual capabilities entering a given boundary and correlates that information with possible undeclared R&D. The machine vision/path analysis engine processes real-time satellite

video feeds to back-calculate vehicular paths entering the boundaries of interest and identifies anomalous paths (see Fig. 1).

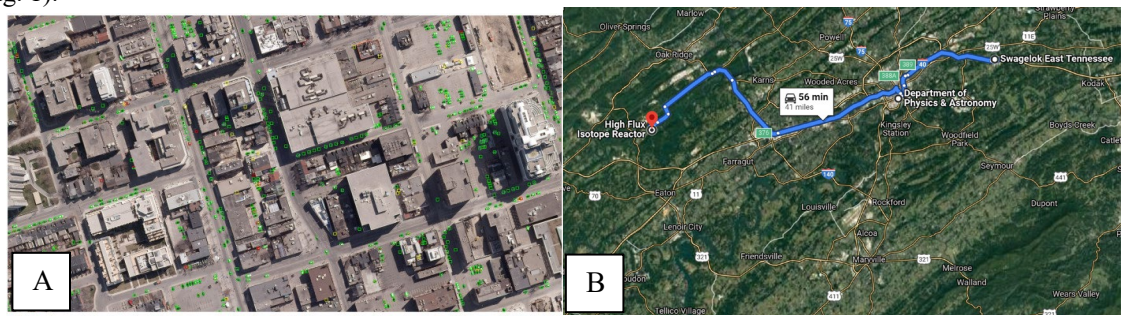

FIG 1. A) The machine vision engine identifies vehicles in the imagery [25]*, and B) the path analysis engine back calculates paths of all vehicles of interest* [26].

Waypoints, or nodes, identified from the anomalous paths seed further analyzes. The geolocation of each anomalous node initiates a workflow using a suite of independent AISs, or engines, which process targeted open-source information streams from public records, social media, and literature. Foundational to the system are the two components tailored solely for IAEA operations: the safeguards neural engine and the open-source platform. The platform has two primary roles: 1) to make specific information requests of the various engines and 2) to properly move and filter data between system components. The safeguards neural engine is the highest level AIS, responsible for consuming data from all of the engines (Fig. 2A), analyzing the data, making actionable recommendations on the likelihood of misuse and diversion, and utilizing user and system feedback to improve the accuracy and confidence in the system (Fig. 2B). On the periphery of the system are the user interface(s) and data archive.

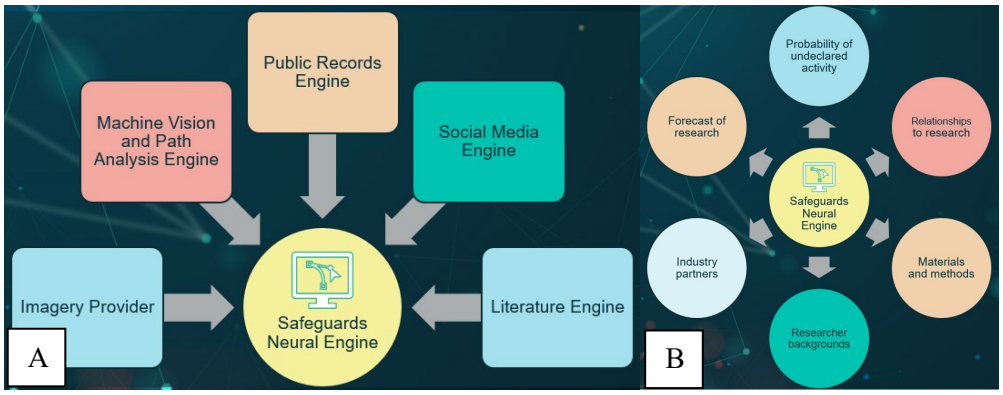

FIG. 2. A) Information feeding into the safeguards neural engine and B) output pathways and analysis.

4. CRITERIA CREATION WORKSHOP

The objective of the Ethical Guidelines for AI Applied to International Safeguards workshop is to explore the relevancy of current ethical AI frameworks to the international safeguards community, elicit the benefits to member states and the IAEA, and understand implementation pathways given current and future structures. Participants engaged in interactive dialog and digital whiteboard activities to provide their insights, thoughts, and concerns regarding data sources, algorithms, bias, transparency, and privacy on the AIS safeguards use case described in section 3.2. The workshop objectives were as follows:

- Provide an overview of AI/ML in the safeguards context
- Introduce ethics criteria and an assessment framework
- Develop consensus on use case impact levels
- Identify safeguards relevant ethical criteria and metrics

The workshop was structured as a moderated virtual collaboration, focused on facilitating an open exchange of ideas on the art of the possible with respect to the future of ethical AI for safeguards. Attendees represented a cross section of the subject matter expertise anticipated to have an impact on AI and ML for

safeguards, contributing to the first-ever study ethical AI for safeguards. This first round of expert elicitation was limited to U.S. Department of Energy (DOE) participants, but future workshops will extend to others.

4.1. Workshop goals and structure

The goal of the workshop was to explore ethical verification criteria for AIS safeguards technologies that will promote ethical integrity in the full life cycle and regulatory processes within this emerging domain. Participants attended a virtual two-part working meeting in September 2022. Each day began with welcoming remarks, a crash course on the collaboration technology[1], and an icebreaker activity fostering an open environment for discussion and ideation. Presentations included an AIS overview, introduction to the ethical AI framework, and introduction to the safeguards use case. Interactive activities involved contemplating the risk level of the use case, culling the pre-populated IEEE goals and requirements on the collaborative digital white board, and conducting an IAEA alignment, barriers, and adoption exercise. At the end of each day, participants provided feedback on the ethical AI framework and the workshop.

4.2. Workshop outcomes

This workshop was the first step in creating a shared vision of ethical AI in safeguards and benefitted from the variety of views and values of those affected by this technology. There were 15 participants representing a mix of IAEA and safeguards experts, data scientists and analysts, and developers. The diversity of expertise, as well as thoughts and comfort levels with AIS, produced incredibly rich and thoughtful discussions throughout the entire workshop.

Initial activities focused on preparing participants for collaborating virtually by getting them comfortable with each other and the AIS use case. The icebreaker activity allowed participants to explore their hopes and fears about applying AI to safeguards; providing participants a chance to open up and become comfortable engaging with the group. The icebreaker revealed broad consensus that AIS would be desirable if it improves the efficiency of conducting verification tasks by reducing analyst time doing repetitive tasks. Similarly, there was considerable concern with AIS tools not being well aligned with the safeguards user community. The use case was introduced in a presentation and followed by a feedback activity examining each of the engines for strengths, additional functionality, deficits, and impact level. In general, participants found the use case realistic and on average of medium impact[2]. They rated the social media and public records engines, for example, as high impact[3] with concerns about releasing personally identifiable information or sensitive personal information, the use of false social media profiles, and confusion with people having the same name.

The primary goal and set of activities for the remainder of both sessions was to select safeguards relevant goals and requirements (discussed in 3.1) as applied to transparency, bias, accountability, and privacy. Direction for participants was to discuss and reflect on the requirements that were deemed most relevant to the IAEA, and those that would be attractive to Member States. A representative outcome of this activity for ethical bias can be seen in Fig. 3, showing just one of the goals for comparison. For illustrative purposes, the goal “System behavior monitoring” was found to be useful for both the IAEA and Member States so it was populated in both columns, but the reasons why it was useful are quite different. For the IAEA, understanding how the system is behaving is important, and there is some nuance involved in determining what is considered a protected characteristic for the IAEA, and how biases around those protected characteristics will be tracked. Member States, on the other hand, might be concerned with misuse of the AIS and their data, and the “System behavior monitoring” goal and requirements may help alleviate those concerns.

---

[1] MURAL is a company with an online platform that connects people with a digital whiteboard used for visual collaboration and exploring ideas.

[2] Medium impact is defined as the anticipated impact to stakeholders is between low and high impacts. Low impact is having little to no anticipated impact on the health, welfare, safety, and ethical values of stakeholders.

[3] High impact is defined as presenting a likelihood of injury or harm to the stakeholders.

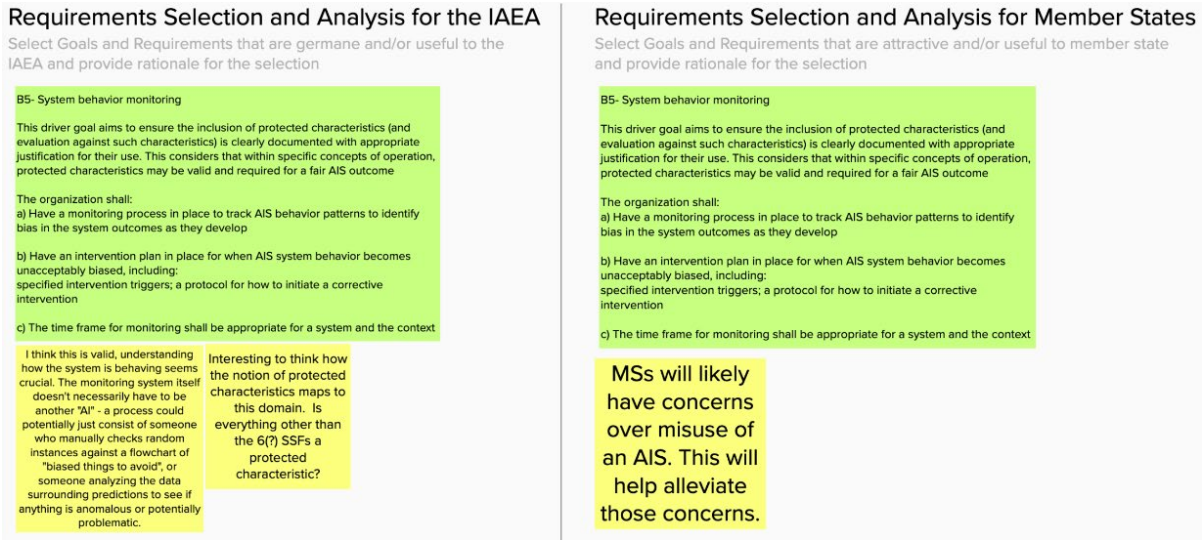

FIG. 3. Snapshot of ethical bias requirements and selection analysis germane to the IAEA (left) and attractive and/or useful to Member States (right) activity.

4.2.1. Requirements selection and analysis for the IAEA

A portion of the IAEA's goals and requirements selection activity, including comments, is shown in Fig. 4. Goals that require clear definitions and communication about the AIS were all selected as relevant and useful to the IAEA (e.g., defining the intended purpose of the AIS, modes of operation, required data inputs and outputs, and user interactions with the AIS). Similarly, goals that increased confidence in system behavior were considered very relevant and useful to the IAEA (e.g., through reporting, auditing and record keeping, human intervention mechanisms to interrogate algorithms, data, and results, and using quality management processes).

Notably, participants found several goals that require quality and security controls to be universal and already standard practice at the IAEA. These already in place structures include ensuring risk mechanisms are in place to decommission systems and software, strong human oversight practices, extensive technical training, privacy controls, and capabilities to identify and respond to bias. Requirements in these categories could provide a useful starting point for implementation as they are well aligned with current IAEA processes and culture.

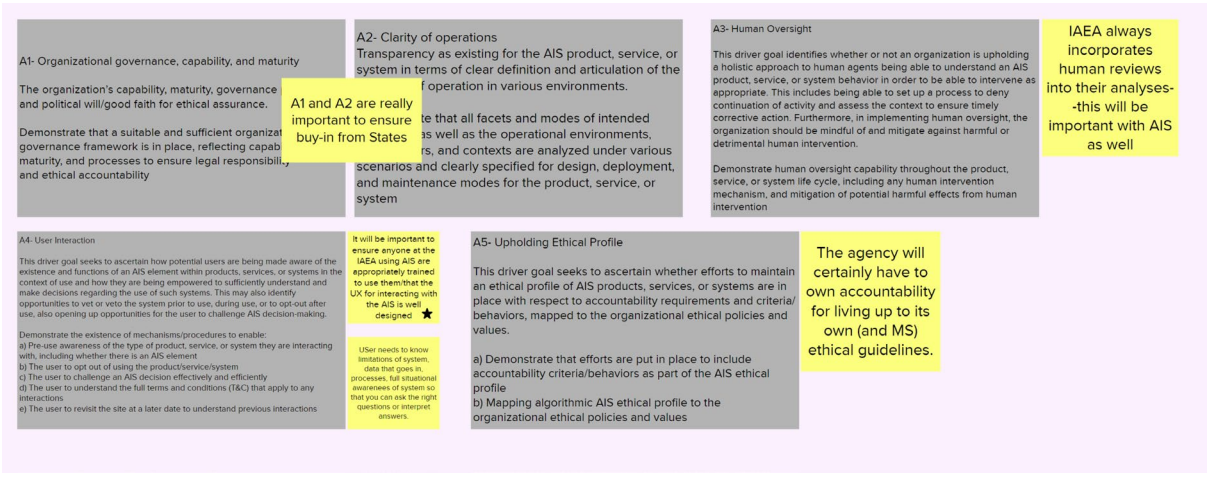

FIG. 4. Snapshot of ethical accountability goals selected as useful to the IAEA

The following details the key considerations about the IAEA that surfaced through the workshop:

— Participants found the *organizational governance* goal and its requirements of having the capability, maturity and governance processes to ensure legal responsibility and ethical assurance important for all key principles. It will be important to develop messaging early about what the AIS is used for and how

transparency, bias, accountability, and privacy will be tracked and addressed. They also thought the requirements and messaging would be important to ensure buy-in from member states.

- Having *clearly defined and communicated concepts and results of operation* were deemed particularly useful for the IAEA to address concerns related to bias and accountability. The clarity and concepts of operation would help ensure continued alignment given the IAEA's high turnover rate.
- Ensuring there is *organizational capability to correct emerging or detected bias* during development, deployment, and operation through risk management, design changes, and compensation mechanisms are useful. The capabilities may help prevent/ address claims from States about discrimination, and they align with existing best practices at the IAEA of justifying decisions that ultimately lead to safeguards conclusions.
- *System behavior monitoring* to track for bias patterns and intervene in a timely manner is also important to understand how the system is behaving, and they recommend a human periodically checks the system for biases.
- Participants thought the *human oversight* goal and requirements for accountability and privacy fit well with existing IAEA practices of incorporating human reviews into analysis. The IAEA would need to have frequent, if not constant, monitoring of any disclosure risk. A human should be able to intervene to prevent or correct any privacy issues.
- Demonstrating efforts are in place that include accountability and transparency criteria and behaviors as part of the *AIS ethical profile* as the IAEA will have to own accountability for living up to its own ethical guidelines.
- The *user interaction* goal seeks to ascertain how potential users are being made aware of the existence and functions of an AIS element within products, services, or systems used for safeguards. The participants found this relevant to the IAEA to ensure there is appropriate training, knowledge of data, processes, and limitations so the user can ask the right questions and interpret answers correctly.
- The *ethical architecture, design, and development for AIS* privacy goal requires organizations to promote a culture of peer and stakeholder accountability, and provides opportunities to raise concerns. This goal and requirement was determined to be relevant for ethical privacy as a fundamental work practice requirement and not just for AIS. Protecting States' confidential data and the IAEA's own work product is a necessary requirement in general.
- *Decommissioning* goal requires having risk and control mechanisms put in place in the decommissioning of AIS. Many of the participants said they have not thought about decommissioning or retirement of AIS before the workshop, but agreed it was an important requirement for privacy. Because of the rapid pace of development of AIS, ethical decommissioning processes may help ensure the privacy of Member State data from cradle to grave. Participants emphasized similar processes are not unique to AIS and should already exist at the IAEA.

4.2.2. Requirements selection and analysis for Member States

The participants also analyzed goals and requirements they considered attractive for Member States. These particular requirements may have the potential to ease Member State concerns about AIS used for safeguards (e.g., being falsely accused of noncompliance, unjustified bias, not knowing how their data will be handled and how privacy will be maintained, and not knowing who is accountable). The primary results are described:

- *Organizational Governance* was an attractive requirement for Member States as it provides assurances that the IAEA is capable of governing, maintaining, operating, and properly decommissioning these AI tools, ensuring information is protected, and that there is no unjustified bias against the Member State.
- *Clarity of AIS Operations* could provide Member States pertinent details on how the AIS will be deployed, and operated.
- Properly executing the requirement of the *System Behaviour Monitoring* goal can mitigate Member States concerns over misuse of an AIS.
- *Justified Use of Protected Characteristics* is critical as the use of a variety of Member State information is likely to be a primary concern, which the IAEA will likely need to assuage. In this context, protected

characteristics could include various forms of proprietary information and safeguards confidential information.

— *End-user Awareness of AIS and Empowerment* seeks to ascertain how potential users, potentially Member States, are being made aware of the existence and functions of an AIS.
— *Ethical Architecture, Design, and Development for AIS* and *Decommissioning* were deemed attractive to Member States as they provide an early understanding of how data and privacy will be handled and maintained throughout the lifecycle of the AIS.

4.2.3. Challenges and further discussion

The following goals and requirements were not selected by the workshop SMEs, however they evoked several questions that warrant additional discussion and consideration for future work:

— Upholding an *ethical profile for AIS* could be useful to maintain ethical accountability at the IAEA.
— Clearly defining how *users interact* with the AIS could be attractive for member states to understand accountability.
— Clearly defining and communicating *context alignment* could be attractive for member states to understand bias and how it is considered in the AIS.
— While likely important to uphold an ethical privacy standard, the *ethical profile* goal caused some confusion about what this means for IAEA.

Participants provided ideas on potential challenges for implementing and validating all of the requirements they selected, a portion of this activity is shown in Fig. 5 for implementing ethical privacy requirements. Some challenges include limited resource allocation, determining the level of information to share with different stakeholders including Member States, figuring out who owns the data before and after its been altered by the AIS, how training data is obtained and where it comes from, and preventing accidental disclosure from externally-facing AIS.

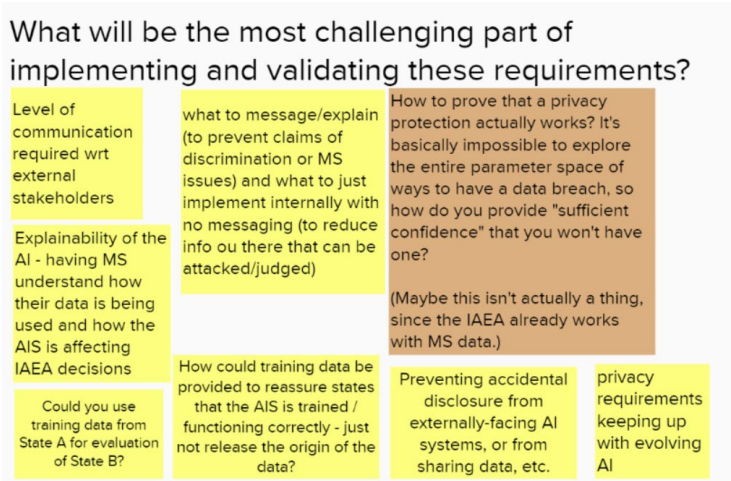

FIG. 5 Snapshot of ethical privacy challenges

4.3. Going forward

In the final activity, participants provided thoughts about their workshop experience and potential next steps; some of the key observations are detailed:

(a) Is this a methodology that could be adopted by the IAEA?
  (i) Yes, but as mentioned a few times, in a stepped approach as the AI develops. Leverage what the IAEA is already doing well and transition it to AIS. Then extend it in the future.
  (ii) Need to bridge the gap between the high-level goals and how they would be implemented by the "boots on the ground." If the implementation becomes too onerous, then the methodology becomes untenable. If the implementation is palatable, then the methodology is great.

(iii) Yes, I think so, and it may encourage the IAEA to revisit transparency on other matters

(b) What portion of this framework do you feel would be most impactful to the IAEA?

(i) Having a governance structure (at least some initial, small-scale framework) would likely help the IAEA. They could probably borrow a lot from the other work being done in the EU on this topic.

(ii) There are many at the IAEA who might not realize that any of this is a thing at all (meaning all of these ethics components of AIS). Simply showing them information like this might provide them with that insight.

(iii) Using it to determine the balance between transparency and privacy.

4.3.4. Potential starting points for implementation

The workshop participants detailed activities that the IAEA has significant depth and experience in that aligns well with the particular Ethical AIS goals. This expertise includes efforts on decommissioning systems, developing privacy controls, and developing and operating an effective and rigorous quality control program (QC). Additionally, participants suggested that cost free experts could be used to help expand the program by integrating appropriate ethical AIS framework requirements into the QC programs.

5. CONCLUSION

The Establishing Ethical Guidelines for Applying AI to IAEA Safeguards project team developed a broad and far-reaching safeguards use case, allowing the team to engage SMEs to analyze and generate a tailored set of ethical considerations around bias, transparency, accountability, and privacy. The first round of elicitation produced pragmatic insights about how an ethical AI framework could be useful for safeguards AIS, what the challengers are, and how ethical requirements could be implemented at the IAEA. It would be beneficial to expand engagement to a broader set of stakeholders including international SMEs, other ethical AI experts, and the IAEA.

While the use case developed covers numerous AIS technologies, it focuses on information that could be gleaned via AIS from outside of a facility to understand the R&D activities occurring within the facility. However, the study does not cover the use of a physical AIS (e.g., robots, drones, etc.) conducting in-field verification tasks. Unmanned AIS capable of navigating and operating within a facility pose unique ethical challenges that will need to be addressed, prior to deployment of such systems. Logical next steps would include 1) validate the practical operation of ethical AIS protocols and procedures with a physical AIS use case and 2) developing an ethical AIS verification system(s) to support safeguards development, advancement, and adoption of AIS tools. Exploring the ethical AI framework using a broader set of stakeholders and at least two different and representative use cases, will provide for a more robust analysis where the ethical considerations can be associated with quantitative and qualitative metrics applicable to the safeguards domain, thus creating a mechanism to verify and validate that the requirements have been met.

Global deployment of AIS and its impact on all of our lives is a certainty, and unethical and sometimes devastating consequences of AIS operation are an unfortunate reality, but they are not inevitable. The IAEA has the potential to reap enormous benefit from AIS by potentially freeing up intellectual resources to focus on critical nuanced analytical challenges. Because of the international stakeholders, diverse staff, and sensitive nature of the data collected and analyzed, the IAEA is well positioned to develop, deploy, operate, and retire AIS in an unbiased, accountable, secure, and transparent manner. A defensible ethical AI framework may support Member State acceptance of AIS adoption particularly if done in a phased approach prioritizing goals and requirements already aligned with existing procedures and processes and building upon the framework as appropriate.

ACKNOWLEDGEMENTS

The Establishing Ethical Guidelines for Applying AI to IAEA Safeguards project team would like to acknowledge U.S. Department of Energy Office of Concepts and Approaches (NA-241) for funding this work, and to thank everyone who participated in the workshop for being open, honest, patient, and excited to create a more responsible and fair future of AI.

REFERENCES

[1] Murphy, C., Barr, J., Establishing Ethical Guidelines for Applying Artificial Intelligence to IAEA Safeguards: Methodology and Use Case Development, internal report, Y-12 National Security Complex, Oak Ridge, 2021.

[2] IEEE SA, Verifying Ethics in AI-based solutions, IEEE SA, 2022.

[3] International Atomic Energy Agency Department of Safeguards, Long-Term R&D Plan, 2012-2023, IAEA, Vienna, Austria, 2013.

[4] International Atomic Energy Agency, Development and Implementation Support Programme for Nuclear Verification 2020-2021, IAEA, Vienna, Austria, 2020.

[5] Jobin, A., Ienca, M., Vayena, E., The global landscape of AI ethics guidelines, Nature Machine Intelligence **1** (2019) 389–399.

[6] West, D. M., The role of corporations in addressing AI's ethical dilemmas, Brookings (2018), https://www.brookings.edu/research/how-to-address-ai-ethical-dilemmas/.

[7] PAI, Partnership on AI (2022), https://www.partnershiponai.org/partners/.

[8] World Economic Forum, Empowering AI Leadership (2022), https://www.weforum.org/projects/ai-board-leadership-toolkit.

[9] U.S. Department of Defense, DOD Adopts Ethical Principles for Artificial Intelligence (2022), https://www.defense.gov/News/Releases/Release/Article/2091996/dod-adopts-ethical-principles-for-artificial-intelligence/.

[10] U.S. Intelligence Community, Principles of Artificial Intelligence Ethics for the Intelligence Community (2020), https://www.dni.gov/files/ODNI/documents/Principles_of_AI_Ethics_for_the_Intelligence_Community.pdf.

[11] U.S. Intelligence Community, Artificial Intelligence Ethics Framework for the Intelligence Community (2020), https://www.dni.gov/files/ODNI/documents/AI_Ethics_Framework_for_the_Intelligence_Community_10.pdf.

[12] Stanford University, Stanford Institute for Human-Centered Artificial Intelligence (2022), https://hai.stanford.edu/.

[13] AI Initiative, AI Initiative (2022), https://aiethicsinitiative.org.

[14] AI Now Institute, AI Now Institute (2022), https://ainowinstitute.org/.

[15] High-Level Expert Group on Artificial Intelligence, Ethics Guidelines for Trustworthy Artificial Intelligence (AI) (2019), https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

[16] Hogenhout, L., A Framework for Ethical AI at the United Nations, United Nations, 2021.

[17] IEEE Standards Association, The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (2022), https://standards.ieee.org/industry-connections/ec/autonomous-systems/.

[18] IEEE Standards Association, The Ethics Certification Program for Autonomous and Intelligent Systems (ECPAIS) (2020) https://standards.ieee.org/industry-connections/ecpais/.

[19] David, S., Goldenstein, J.-C., Hessami, A., Jordan, S., Shaw, P., Watson, N., Weger, G., IEEE Use Case–Criteria for Addressing Ethical Challenges in Transparency, Accountability, and Privacy of CTA/CTT, IEEE, 2020.

[20] Business Wire, City of Vienna Earns IEEE AI Ethics Certification Mark; Reinforcing Commitment to Digital Humanism Strategy (2021), https://www.businesswire.com/news/home/20211115005200/en/City-of-Vienna-Earns-IEEE-AI-Ethics-Certification-Mark-Reinforcing-Commitment-to-Digital-Humanism-Strategy.

[21] IEEE Standards Association, IEEE CertifAIEd™ – Ontological Specification for Ethical Transparency, The Institute of Electrical and Electronics Engineers, Inc., 2022.

[22] IEEE Standards Association , IEEE CertifAIEd™ – Ontological Specification for Ethical Accountability, The Institute of Electrical and Electronics Engineers, Inc., 2022.

[23] IEEE Standards Association, IEEE CertifAIEd™ – Ontological Specification for Ethical Bias, The Institute of Electrical and Electronics Engineers, Inc., 2022.

[24] IEEE Standards Association, IEEE CertifAIEd™ – Ontological Specification for Ethical Privacy, The Institute of Electrical and Electronics Engineers, Inc., 2022.

[25] C. R. D. C. Stuparu DG, Vehicle Detection in Overhead Satellite Images Using a One-Stage Object Detection Model., Sensors (Basel) **20** 22 (2020)

[26] Google, High Flux Isotope Reactor to Swagelok East Tennessee (2022), https://www.google.com/maps/dir/High+Flux+Isotope+Reactor,+Oak+Ridge,+TN+37830/Department+of+Physics+%26+Astronomy,+Circle+Drive,+Knoxville,+TN/Swagelok+East+Tennessee,+Blakely+Court,+Knoxville,+TN/@35.9629712,-84.3199174,70080m/data=!3m2!1e3!4b1!4m20!4m1. [Accessed 11 10 2021].

DISCLAIMER

COPYRIGHT NOTICE