CBWNet

Strengthening the norms against chemical and biological weapons

WORKING PAPER No. 20, September 2026

Artificial Intelligence Assurance for the OPCW: Governing Trustworthy AI in CWC Implementation

Alexander Kelle

CBWNet

Artificial Intelligence Assurance for the OPCW: Governing Trustworthy AI in CWC Implementation

Alexander Kelle
Institute for Peace Research and Security Policy at the University of Hamburg (IFSH)

Executive Summary
Artificial intelligence (AI) is a cross-cutting condition of contemporary organisational practice and is already relevant to the Organisation for the Prohibition of Chemical Weapons (OPCW). It affects routine functions such as drafting, translation, knowledge retrieval, and information management, as well as potentially consequential functions connected to declarations processing, pattern recognition, chemical detection and identification, data fusion, simulation-based training, and monitoring of scientific and technological developments. This means that the OPCW cannot treat AI solely as a discrete technical project. It must adapt to an AI-enabled operating environment while differentiating ordinary productivity tasks from systems that can affect core functions of CWC implementation.

This paper argues that AI used in, or materially supporting, core treaty-implementation functions should be governed through an organisation-wide AI assurance architecture built around a continuing lifecycle process for ensuring that an AI-enabled system is trustworthy for a defined task and operating context. The framework combines hierarchical assurance, which provides formal controls through risk classification, defined requirements, validation, quality assurance and control, audit trails, secure data governance, review processes, and incident procedures with relational assurance to build confidence through participation, explanation, contestability, user training, feedback, and co-design with those who operate, oversee, or are affected by AI-enabled systems.

These approaches are applied through six OPCW-specific criteria: CWC alignment and defined purpose; confidentiality and data governance; human authority; validated performance and security; accountability and auditability; and equity and participation. The paper applies the framework across four areas: organisational governance; verification, attribution, and analytical integrity; assistance and protection; and international cooperation and capacity building. It proposes lifecycle oversight, secure infrastructure, assurance-sensitive procurement, defined responsibilities, and a small AI Assurance and Data Governance Unit or cross-functional mechanism building on the Technical Secretariat's emerging AI capacity.

AI-enabled systems that materially inform core CWC functions or could significantly affect evidentiary, operational, confidentiality, or States Party interests should be treated as high-assurance systems. This normally includes declaration analysis, anomaly detection, data fusion, mission planning, chemical identification, and operational assistance. Outputs should serve as decision support in documented, human-led processes, not autonomously make compliance, attribution, triage, treatment-allocation, or other consequential determinations.

Implementation should be phased. Initially, the Technical Secretariat should map and risk-classify AI uses, establish safeguards for confidential and operationally sensitive information, and undertake a limited number of bounded, controlled pilots. It should then adopt an AI Assurance and Data Governance Policy; integrate assurance into procurement, project approval, information-security, and audit processes; develop domain-specific procedures; and build staff and States Party capacity. Over time, AI assurance should become a routine element of OPCW governance, supported by SAB advice, monitoring of relevant scientific and technological developments, reporting to policy-making organs, and inclusive cooperation among States Parties.

1 Introduction

Artificial intelligence (AI) is a crosscutting governance issue for implementation of the Chemical Weapons Convention (CWC) and for the routine functioning of the Organisation for the Prohibition of Chemical Weapons (OPCW). Its relevance extends from everyday internal activities – such as drafting, translation, information retrieval, knowledge management, and administrative support – to verification and investigation-related analysis, chemical detection and identification, assistance and protection, training, and international cooperation. In the CWC's post-destruction phase, preventing the re-emergence of chemical weapons, supporting compliance assessment, and responding to allegations of use have assumed greater prominence.[1] These tasks require the OPCW to handle complex and heterogeneous information while maintaining confidentiality, evidentiary integrity, political impartiality, and confidence among States Parties.

AI may assist in meeting these demands. The Scientific Advisory Board's (SAB) Temporary Working Group (TWG) on Artificial Intelligence identifies potential applications in declarations processing, anomaly and pattern detection, spectral and sensor-data interpretation, chemical detection and identification, data fusion, simulation-based training, secure knowledge management, and monitoring relevant scientific and technological developments.[2] AI-enabled synthesis planning, property prediction, automated laboratories, and remote experimentation do not create an entirely new category of dual-use concern: related scientific and technical activities pre-date the present AI wave. They may, however, lower informational and design barriers, accelerate the iteration of technical options, and complicate the interpretation of indicators that have traditionally been associated with specialised infrastructure and expertise. The resulting technical challenge is to assess how AI may alter the relationship between information, physical resources, technical skills, and observable indicators of chemical activity.[3]

The central organizational question is how the OPCW can use AI in a way that is consistent with the Convention's object and purpose, the stringent requirements of the Confidentiality Annex, and the procedural standards that underpin credible and even-handed implementation.[4] Opaque or weakly validated systems may compromise the defensibility of compliance-relevant analysis. Insecure deployment could expose sensitive state and industry information. Automation may distort professional judgement or blur responsibility, while unequal access to technical expertise, data, and capacity to contest AI-supported decisions could deepen disparities among States Parties. AI governance is therefore a requirement of treaty implementation, rather than a discretionary matter of technological modernisation.

This paper argues that the appropriate response is an organisation-wide AI assurance architecture. AI assurance is understood here as a continuing lifecycle process through which the OPCW defines a clearly circumscribed use case, assesses context-specific legal, operational, security, and policy risks, establishes evidence that the system is fit for its authorised purpose; and reviews that conclusion as the system, data, or operating environment changes. It is not a one-off technical test or procurement exercise. Rather, it provides a means of enabling useful applications while ensuring that human expertise, legal responsibility, and multilateral legitimacy remain central to CWC implementation.

[1] Alexander Kelle, "The CWC at 25: from verification of chemical weapons destruction to attribution of their use," *The Nonproliferation Review*, 28 (4-6), 2021, pp.319-336.

[2] OPCW, *Artificial Intelligence. Report of the Scientific Advisory Board's Temporary Working Group*, document SAB/REP/1/26, The Hague, March 2026.

[3] OPCW, Report by the Director-General: Response to the Report of the Scientific Advisory Board on Developments in Science and Technology to the Fifth Special Session of the Conference of the States Parties to Review the Operation of the Chemical Weapons Convention, document RC-5/DG.2, The Hague, 22 February 2023.

[4] OPCW, *Convention on the Prohibition of the Development, Production, Stockpiling and Use of Chemical Weapons and on Their Destruction (CWC)*, consolidated text, incorporating changes effective 7 June 2020.

The proposed approach combines *hierarchical assurance* and *relational assurance*. The former comprises formal, evidence-based controls: risk classification, assurance plans, technical and security validation, data and system documentation, audit trails, governance gates, and incident procedures. The latter concerns the confidence-building relationships surrounding an AI system: participation, appropriate explanation, contestability, user training, feedback, and capacity building for inspectors, analysts, designated laboratories, National Authorities, States Parties, and cooperation partners. Formal controls are indispensable, but they are insufficient on their own in an organisation whose authority rests on both professional competence and continuing state party confidence in the organisation's competence and integrity.

The framework applies these forms of assurance through six criteria: CWC alignment and defined purpose, confidentiality and data governance, human authority, validated performance and security, accountability and auditability, and equity and participation. These criteria adapt general AI-governance principles to the distinctive legal, institutional, and political context of the OPCW.

The paper has three purposes. First, it conceptualises AI assurance for CWC implementation, connecting lifecycle risk management to established arms-control concerns, most notably verification confidence, analytical integrity, chemical safety and security, confidentiality, accountability, and fairness. Second, it identifies assurance requirements across four implementation areas: organisational governance; verification, attribution, and analytical integrity; assistance and protection; and international cooperation and capacity building. Third, it advances a sequenced implementation path through which the OPCW can translate the TWG's agenda into accountable institutional practice.

The following section situates AI within the changing CWC environment and reviews the OPCW's emerging institutional agenda. Section 3 develops the assurance concept and the six crosscutting criteria. Section 4 considers how the TWG's findings contain the nucleus of an assurance agenda. Sections 5 to 8 apply the framework to organisational governance, verification and analytical integrity, assistance and protection, and capacity building and international cooperation. The final section sets out a phased roadmap for embedding AI assurance in routine OPCW governance.

2 The changing CWC environment and the OPCW's emerging AI agenda

Artificial intelligence has become more salient for the CWC because the treaty now operates in a post-CW destruction environment. Verification of declared stockpile elimination is no longer its principal operational focus. Instead, preventing the re-emergence of chemical weapons, supporting compliance assessment, and addressing alleged use have become more prominent.[5] These tasks require the OPCW to assess heterogeneous information, maintain the credibility of its findings under political scrutiny, and adapt to scientific and technological (S&T) change without compromising confidentiality or impartiality. AI-assisted data processing, pattern recognition, and analytical support may help meet these demands, but their use must remain compatible with the Convention's evidentiary, confidentiality, and non-intrusion requirements.

The relevance of AI will also vary across CWC functions and among States Parties. National Authorities and States Parties may develop or procure their own tools for preparing declarations, managing national implementation information, or supporting chemical-safety and security activities. They may be reluctant to place sensitive information in systems provided by the Technical Secretariat (TS) or external suppliers. This reinforces the need for clarity about the Secretariat's own authorised systems, their limitations, the data that may be processed, and the forms of independent evaluation and consultation through which States Parties can have confidence in their use.

[5] Alexander Kelle, "The CWC at 25."

The OPCW's AI agenda has developed through a progression from longer-standing scientific monitoring to more structured institutional engagement. SAB consideration of relevant scientific and technological developments – including AI-assisted chemistry, data-intensive analysis, and related issues – can be traced through its work from approximately the mid-2010s to the Fifth Review Conference in 2023.[6] In 2024, the Organisation launched its AI Research Challenge and convened the Global Conference on the Role of Artificial Intelligence in Advancing the Implementation of the Chemical Weapons Convention in Morocco.[7] It then established the SAB's TWG on Artificial Intelligence in January 2025. The June 2025 Shanghai workshop on AI and chemical safety and security management, together with the Advisory Board on Education and Outreach's consideration of AI governance and ethics, further broadened the Organisation's institutional engagement.[8]

The TWG's March 2026 report provides the most comprehensive basis for the next phase of work. Its recommendations address continued monitoring, engagement with scientific, technical, and industry communities, AI misuse indicators, human-supervised declarations processing and verification support, secure computing environments, detection and identification tools, AI-supported training, and capacity building and twinning for States Parties.[9] While important in agenda-setting terms, the TWG report does not address the need to translate clearly circumscribed AI use cases into a coherent governance framework based on. The following sections develop that framework as AI assurance for the OPCW. In translating this agenda into practice, care needs to be taken to distinguish treaty-linked operational assurance from broader debates about national AI governance in CWC States Parties. Where a proposed Secretariat capability may affect inspection practice, data-sharing expectations, or matters on which States Parties hold different legal and policy positions, consultation and, where necessary, direction from the relevant policy-making organs may be required.

3 AI assurance and cross-cutting criteria for the OPCW context

AI assurance is a lifecycle process for measuring, evaluating, and communicating whether an AI-enabled system is trustworthy for a defined purpose and operating environment. Although the assurance approach and criteria proposed here can guide the deployment of other new technologies and scientific methods in the OPCW context, AI warrants particular attention because of the rapid evolution of its capabilities. It is a governance architecture that must connect substantive norms, risk classification, technical evidence, organisational authority, external oversight, and feedback from real-world operation across the full life of an AI-enabled capability. In the OPCW context, it means establishing evidence that an AI system supports CWC implementation within acceptable legal, security, operational, and policy parameters. AI assurance begins with the proposal of a clearly circumscribed AI use case and continues through design, procurement or in-house development, validation, deployment, monitoring, modification, and

6 See e.g. OPCW, Report by the Director-General: Report of the Scientific Advisory Board on Developments in Science and Technology to the Fifth Special Session of the Conference of the States Parties to Review the Operation of the Chemical Weapons Convention, document RC-5/DG.1, The Hague, 22 February 2023.

7 OPCW, Note by the Technical Secretariat: Global Conference: The Role of Artificial Intelligence in Advancing the Implementation of the Chemical Weapons Convention, document S/2299/2024, The Hague, 25 June 2024; OPCW, Note by the Technical Secretariat: The OPCW Artificial Intelligence Research Challenge, document S/2301/2024, The Hague, 4 July 2024.

8 OPCW, Note by the Technical Secretariat: Invitation to Apply for a Workshop on Artificial Intelligence and Chemical Safety and Security Management, Shanghai, China, 17 – 20 June 2025, document S/2373/2025, The Hague, 25 February 2025; OPCW, Report of the Nineteenth Session of the Advisory Board on Education and Outreach, document ABEO-19/1, The Hague, 26 June 2025, p.6f.

9 OPCW, Artificial Intelligence. Report of the Scientific Advisory Board's Temporary Working Group, document SAB/REP/1/26, The Hague, March 2026.

retirement. It is therefore not a one-time test or procurement check, but a continuing process through which the Organisation determines whether an AI system is and remains fit for its authorised function.[10]

The framework should combine hierarchical and relational assurance. Hierarchical assurance provides formal, evidence-based controls: risk classification, defined use requirements, data-quality and system-quality controls, assurance plans, documentation, testing and validation protocols, audit trails, and incident procedures. It establishes whether a system meets the requirements defined for its authorised function and whether residual risks are acceptable in the relevant operating context. Relational assurance addresses the confidence-building relationships around an AI system. It requires appropriate participation, explanation, contestability, feedback, and capacity building for those who operate, oversee, or are affected by AI.[11] In the OPCW context, this includes its policy-making organs, TS staff, including inspectors, designated laboratories, national authorities, other relevant actors in States Parties, and cooperation partners.

The AI assurance criteria below do not imply that all technical specifications can, or should, be negotiated in detail by States Parties. The TS should develop task-specific assurance requirements within its operational competence, drawing on existing CWC procedures, laboratory-quality practices, information-security rules, procurement requirements, and applicable professional standards. The cross-functional coordination mechanism proposed below (see section 5) should maintain templates, advise on proportionate requirements, and verify that the relevant business owner has addressed them. For high-impact functions or proposed changes with significant policy implications, the Secretariat should seek appropriate advice and consult States Parties through established channels.

Criterion	Core question for AI assurance
CWC alignment and defined purpose	Does the proposed use have a clear treaty-linked function, remain necessary and proportionate to that function, and respect the respective responsibilities of the Technical Secretariat and policy-making organs?
Confidentiality and data governance	Are data, models, access arrangements, testing environments, retention practices, and supplier arrangements compatible with Article VIII(5), Article VI(10), the Confidentiality Annex, and the Confidentiality Policy adopted by the OPCW?
Human authority	Are consequential assessments and decisions made by authorised officials, with users able to question, override, and escalate AI outputs?
Validated performance and security	Has the system been tested and validated against the relevant task, conditions, uncertainty, errors, manipulation risks, and operational constraints?
Accountability and auditability	Are ownership, approval, documentation, logging and review (for traceability), incident reporting, and suspension responsibilities clearly allocated?
Equity and participation	Can States Parties and operational users meaningfully understand, assess, and where appropriate contest AI-enabled practices, without creating avoidable disparities in access or governance capacity?

[10] Douglas Robbins, Ozgur Eris, Ariel Kapusta, Lashon Booker, and Paul Ward, AI ASSURANCE. A Repeatable Process for Assuring AI-enabled Systems, June 2024, McLean, VA: MITRE; UK Department for Science, Innovation & Technology, *Guidance. Introduction to AI assurance*, 12 February 2024.

[11] Yumeng Guo, Jonathan Foster, Alastair R Buckley, and Georgios Zoukas, "Assuring trustworthy AI: An integrative review of hierarchical and relational assurance approaches," *International Journal of Information Management Data Insights*, 6 (2026), 100415, p.5.

These criteria provide the common analytical lens for the remainder of the paper, but their practical priority varies by implementation pillar. In verification, treaty purpose, confidentiality, controlled access to information, data quality, evidentiary integrity, validation, and reviewability are all central. In assistance and protection, priority attaches to operational reliability, human responsibility, resilience, and secure handling of sensitive information. In international cooperation, accessibility, shared validation, meaningful participation, and sustainable local capability should be prioritised. Across all pillars, quality assurance and quality control must apply both to data entering a system and to the system's performance, documentation, security, and use in practice.

The criteria should guide AI system-specific assurance plans rather than operate as a universal checklist applied mechanically to every tool. High-assurance systems are AI-enabled systems whose outputs materially inform core CWC implementation functions or could significantly affect evidentiary, operational, confidentiality, or State Party interests; they require enhanced, documented evidence of treaty alignment, secure design, task-specific validation, substantive human review, traceability, and periodic reassessment. Applications affecting declarations, inspection preparation, analytical conclusions, chemical identification, or operational assistance will normally fall within this category. Lower-risk tools may be subject to proportionate controls, but all systems must remain within approved functions and confidentiality safeguards.

4 Temporary Working Group findings and AI assurance implications

The SAB TWG on Artificial Intelligence provides the first comprehensive assessment from within the OPCW of how AI is reshaping chemistry, verification, and chemical security.[12] It treats AI as a cross-cutting enabling technology that already affects multiple aspects of the Organisation's work and therefore requires sustained institutional attention rather than ad hoc experimentation.

The TWG identifies five AI application clusters of particular relevance to the OPCW. First, AI-assisted synthesis planning and property or toxicity prediction may help anticipate novel chemicals of concern and support safer peaceful chemistry, but they also have a dual-use character because they may lower barriers to harmful design.[13] Second, AI can improve interpretation of complex spectral and sensor data generated by OPCW-relevant analytical systems, including under field conditions where rapid identification matters.[14] Projects under the OPCW AI Research Challenge illustrate this, exploring different tools for enhancing forensic and detection capabilities.[15] Third, the convergence of AI with automated and remote laboratories may alter observable indicators of chemical activity, requiring the OPCW to reassess verification approaches that were designed around more traditional infrastructures.[16] Fourth, AI may strengthen data fusion and information analysis by helping experts organise large declaration datasets, detect anomalies, and integrate open-source or remote-sensing information into broader analytical workflows.[17] Fifth, the report highlights AI-enhanced training and simulation, especially extended-reality environments and AI-supported scenario generation for inspectors and responders.[18]

[12] "OPCW releases landmark report on AI and the Chemical Weapons Convention," The Hague: OPCW, 11 March 2026.

[13] OPCW, *Report of the Scientific Advisory Board's Temporary Working Group on Artificial Intelligence*, document SAB/REP/1/26, The Hague: OPCW, 3 March 2026, pp.71-73.

[14] OPCW, *Report on Artificial Intelligence*, pp.73-77.

[15] "OPCW AI Research Challenge: Harnessing AI tools to enhance global chemical security", The Hague: OPCW, 25 April 2025. See also Agenda Item 14 on the results of the OPCW AI research challenge in OPCW, *Report of the Scientific Advisory Board at its Fortieth Session*, The Hague: OPCW, 8 May 2026, pp.16-23.

[16] OPCW, *Report on Artificial Intelligence*, pp.30-34.

[17] OPCW, *Report on Artificial Intelligence*, pp.77-79.

[18] OPCW, *Report on Artificial Intelligence*, pp.87-98.

Although the TWG does not use the language of assurance explicitly, several of its recommendations point in that direction. The first is continuous monitoring of AI developments relevant to chemistry, data analysis, and changing indicators of chemical activity, ideally embedded in existing science and technology review processes and supported by dedicated expert structures when needed.[19] A second concerns human-in-the-loop use of AI in declarations processing and verification support, including digitisation, structuring, anomaly detection, and pattern recognition, while preserving analyst and inspector responsibility for compliance-related judgments. [20]

The report also presents secure computing environments and data governance as prerequisites for AI use in OPCW contexts. Given the sensitivity of declarations, inspection findings, and industrial information, it calls for robust information-security measures, including secure and compartmentalised networks, strict access controls, and scrutiny of externally procured systems for compatibility with OPCW confidentiality requirements. In addition, the TWG recommends developing indicators of misuse of AI in chemistry through consultation with States Parties and external experts, so that suspicious behavioural or technical patterns can inform both national governance and OPCW analytical practice. [21] Finally, it proposes capacity-building and twinning initiatives to address uneven AI-related capabilities among States Parties, including training on AI tools, misuse indicators, and governance options, as well as structured peer learning between more and less advanced national contexts. [22]

Together, the TWG's recommendations supply the operational building blocks of the six-criterion assurance framework developed here: risk differentiation, human authority over consequential judgements, confidentiality-compatible design, validation and traceability, and capacity building for meaningful participation. Such a framework should also be compatible with widespread ethical guidelines for AI governance.[23] However, in contrast to generic guidelines the AI framework would not remain at the level of aspirational principles. Instead, it would translate principles such as accountability into audit trails and role allocation within the TS, safety and robustness into validation and secure deployment, transparency into explainability in a controlled environment, confidentiality into protected data governance, and fairness into design choices for AI systems that account for capacity limitations and contestability requirements in CWC implementation.

5 Organisational governance as a core component of AI assurance

AI assurance requires organisational arrangements that enable the OPCW to authorise, secure, evaluate, and review AI-enabled systems used in CWC-related work. AI should therefore be governed as an institutional capability, not acquired as a series of stand-alone technology projects. AI use by the TS requires clear ownership, defined review procedures, secure technical infrastructure, and the capacity to assess whether systems remain appropriate for their authorised functions.[24]

19 OPCW, *Report on Artificial Intelligence*, p. 25.

20 OPCW, *Report on Artificial Intelligence*, p. 50.

21 OPCW, *Report on Artificial Intelligence*, p. 42.

22 OPCW, *Report on Artificial Intelligence*, p.65.

23 Jessica Fjeld et al., *Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI*, Berkman Klein Center Research Publication No. 2020-1, January 2020; Nicholas Kluge Corrêa et al., "Worldwide AI Ethics: A Review of 200 Guidelines and Recommendations for AI Governance," *Patterns* 4, No. 10, 2023: 100857, https://doi.org/10.1016/j.patter.2023.100857..

24 Some of these elements are contained in the TWG AI report.

A secure internal assistant illustrates this requirement. As envisaged by the SAB TWG on AI, such a tool could operate in an air-gapped or otherwise protected environment to support authorised translation, summarisation, drafting, answering questions, and access to technical information. Its value is not simply efficiency: it provides a controlled alternative to public AI services for routine tasks involving sensitive material. In line with CWC confidentiality requirements, access controls, usage logging, staff guidance, and auditable records should be built into the system from the outset, and protected operational information should not be used for model training or transferred to external providers without explicit authorisation.[25]

Procurement is a principal assurance lever. AI-related specifications and contracts should require interoperability with OPCW systems; secure hosting and access controls; system, model, and data documentation; task-relevant performance evidence; logging and audit capabilities; and arrangements for human review. For high-assurance systems, contracts should also require notification of material changes, incident cooperation, defined retention and deletion practices, and an OPCW right to suspend use if assurance conditions are no longer met. These safeguards preserve the Secretariat's ability to assess and control externally supplied systems rather than becoming dependent on opaque vendor claims or inflexible technical architectures.[26]

Organisational roles should be allocated across existing functions. The relevant business division should own each approved use case and remain responsible for its operational use. IT and information-security functions should oversee secure architecture and access; legal, confidentiality, and data-governance functions should advise on applicable constraints; procurement should embed assurance requirements in contracting; and internal audit should review compliance and control effectiveness. The SAB should continue to provide independent scientific and technical advice, including on emerging technologies, validation approaches, chemical-model risks, and assurance criteria. To strengthen the usefulness of advice on AI-related questions, the OPCW could seek an appropriate balance of expertise that includes AI developers, cybersecurity and data-governance specialists, alongside the more traditional chemical and other scientific expertise required for CWC advice.

A practical coordinating mechanism should build, where possible, on existing Secretariat capacity, including the knowledge-management and AI function recently created in the TS. It could take the form of a small AI Assurance and Data Governance Unit, supported by a cross-functional review group convened for high-assurance systems. Its purpose would be coordination rather than independent operational control. The mechanism should neither replace domain responsibility in TS divisions nor assign the SAB routine membership or operational responsibility.[27] Where appropriate, it may request ad hoc advice from the SAB, the Advisory Board on Education and Outreach (ABEO), external specialists, or States Parties through established processes.

The coordinating mechanism should support a lifecycle process comprising (1) initiation and design, (2) development or procurement, (3) deployment decision, and (4) post-deployment monitoring and review of AI systems. At initiation, a short concept note should identify the purpose, users, data sources, expected benefits, treaty function, and foreseeable risks. Approval to proceed should be conditional on an AI Assurance Plan specifying performance expectations, confidentiality and security controls, testing methods, documentation requirements, user responsibilities, and thresholds for residual risk.[28] Assurance must also address operational continuity. A system may

[25] See the TWG on AI recommendation 13 on secure internal chatbots or assistants, in *Final Report*, pp. 15, 49.

[26] Merve Hickok, "Public procurement of artificial intelligence systems: new risks and future proofing," in *AI & Society,* 39, 1213–1227 (2024); OECD, "AI in public procurement," in *Governing with Artificial Intelligence: The State of Play and Way Forward in Core Government Functions*, OECD Publishing, Paris, 2025, pp.204-211.

[27] Terminology used for such a panel or mechanism is secondary. Essential is its cross-functional nature to review, challenge, approve, escalate, and periodically reassess AI across its lifecycle.

[28] National Institute of Standards and Technology, *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*, Gaithersburg, MD: NIST, 2023.

become unsuitable because its performance degrades, its model or data environment changes, a supplier alters the product, a security concern arises, or the task itself evolves. High-assurance systems should therefore have manual procedures, alternative technical arrangements, or other redundancy proportionate to the operational consequence of withdrawal.

During development or procurement, assurance activities should be incorporated into specifications and work plans rather than added after deployment. They should include appropriate documentation, security assessment, controlled user testing, and task-specific validation. Any operational use should require a documented deployment decision by the responsible Secretariat unit, informed by the assurance material and, for significant systems, by cross-functional review. Systems affecting inspection planning, compliance-related assessment or other functions with possible policy implications may also require appropriate consultation with States Parties or direction by the relevant policy-making organs.

Once deployed, significant systems should have a designated operational owner, relevant performance indicators, error and incident logs, and defined triggers for reassessment. Material changes in the model, data environment, operating context, or intended use should prompt review before continued deployment. Periodic assessment should determine whether the system remains fit for purpose, requires modification or restriction, or should be retired. This structure makes assurance a routine part of organisational governance rather than a one-time approval exercise.[29]

6 Verification, attribution, and analytical integrity

Verification-related AI applications should be assessed by their contribution to analytical integrity, i.e. whether they improve the TS's ability to produce reliable, reviewable, and treaty-relevant findings. Relevant applications include declarations processing, anomaly detection, data fusion, mission preparation, chemical detection and identification, and support for alleged-use investigations. Because these tools can influence the evidentiary chain underlying verification activity, their assurance requirements should be calibrated to the specific CWC function and operational consequence of each use case. In this area, assurance must also be responsive to the limits of available data and to the differentiated legal and political setting of CWC verification. Information that may be relevant to one context is not necessarily authorised for use in another; remote-sensing or other externally derived information, for example, should be used only where authorised through applicable procedures and policy direction. This approach is also reflected in the broader debate on 'AI verification', where a suite of mechanisms has been proposed that underline the need to embed AI tools in treaty-compatible inspection and oversight practices.[30]

6.1. Declarations and anomaly detection

The TWG recommends digitising historical and current declarations, extracting and structuring relevant information, and maintaining it in a contestable environment. AI may assist comparison across large and heterogeneous records, identification of inconsistencies, and prioritisation of material for analyst review. For this use case, assurance should focus on preserving the fidelity of original submissions; documenting transformations, classifications, and system outputs; defining authorised data access and reuse; and establishing error-correction procedures where extracted or structured data prove inaccurate. Comparable concerns have arisen in proposals for 'arms

[29] Carolin Gotsch and Jörg Puchan, "Harmonising innovation and governance: A lifecycle model for high-risk AI systems under the European AI Act," in Markus Böhm and Jürgen Wunderlich (Hrsg.): *AKWI Jahrestagung 2024, Lecture Notes in Informatics (LNI)*, Bonn: Gesellschaft für Informatik, 2024, pp.79-95; United Nations System Chief Executives Board for Coordination, *Principles for the Ethical Use of Artificial Intelligence in the United Nations System*, New York: United Nations, 2022.

[30] Matthew Mittelsteadt, *AI Verification: Mechanisms to Ensure AI Arms Control Compliance*," Washington, D.C.: Center for Security and Emerging Technology, February 2021.

control for artificial intelligence', which emphasise test datasets, retraining checks, and behavioural triggers as practical means of verifying compliance with agreed constraints on AI-enabled military systems.[31]

The same reasoning applies to anomaly detection and pattern recognition. Such tools may help analysts identify inconsistencies, trends, or records meriting further inquiry, particularly where information must be compared across time, facilities, languages, and reporting formats. Yet an anomaly is not evidence of non-compliance. A model may be reacting to uneven data quality, translation variation, changing national reporting practices, or an unrepresentative training set. Experience from nuclear safeguards suggests that AI models used for surveillance and anomaly detection can misclassify events when trained on unrepresentative data, reinforcing the need for domain-specific validation under realistic operating conditions.[32] The problem is especially acute for declarations under the CWC. States Parties differ substantially in the kinds, number, history, and distribution of declarable facilities and activities, while reporting formats, languages, and prior clarification practices vary over time. An apparent anomaly may therefore reflect the normal profile of a particular State Party or facility category rather than a matter warranting concern. AI system outputs should consequently be treated as contestable inputs into an assessment process, not as determinations or automatic triggers for adverse action. Before use, systems should be tested across a variety of declaration types and facility profiles, documentation should specify thresholds, assumptions, error rates, uncertainty, and known failure modes, and analysts should record the basis on which they accept or reject an AI system input into their decision-making.

6.2. Data fusion and mission planning

AI-supported data-fusion and mission-planning systems are high-assurance applications because they may shape the operational focus of inspections and thereby affect the evidentiary environment in which later judgements are made. Their evaluation should extend beyond average technical accuracy to robustness under noisy, incomplete, manipulated, or politically contested information. The system's authorised function must be explicit: it may assist logistics, contextual review, or prioritisation, but should not autonomously select inspection targets, determine the scope of access, or reach compliance conclusions. It should also recognise the practical diversity of inspected facilities and inspection practice: inspection reports and facility descriptions reflect both the distinctive characteristics of each site and the professional judgement of individual inspectors. AI may usefully identify recurring questions, patterns, or points requiring preparation, but it cannot replace context-sensitive inspection judgement. Analyses of AI/ML in nuclear safeguards similarly warn that biased datasets, manipulated information, and over-reliance on automated analysis can erode confidence in verification.[33]

Under CWC Article IX, a challenge inspection serves the sole purpose of clarifying and resolving questions of possible non-compliance; requests must remain within the Convention's scope and be supported by appropriate information. An AI output cannot itself satisfy those legal predicates. This point is particularly important because there have been no challenge inspections or investigations of alleged use conducted under the CWC procedures from which a large, directly comparable training corpus could be derived. Any future exploration of AI support for non-routine contexts would therefore require exceptional caution, clear legal and forensic standards, and a careful assessment of whether lessons from the Fact-Finding Mission, Declaration Assessment Team, Investigation and Identification Team, or other relevant practice are transferable to the distinct

[31] Thomas Reinhold, "Arms Control for Artificial Intelligence," in Thomas Reinhold and Niklas Schörnig (eds.), *Armament, Arms Control and Artificial Intelligence. The Janus-faced Nature of Machine Learning in the Military Realm*, Cham: Springer, 2022, pp.211-226.

[32] Hyuk Kim, "Artificial Intelligence and Its Convergence with Nuclear Safety, Security and Safeguards," Open Nuclear Network, 13 June 2024.

[33] Francisco Parada Iturria et.al., "AI for Nuclear Safeguards Verification," Oak Ridge National Laboratory, 2024.

legal purposes of the CWC. AI may inform authorised expert assessment, but only within a documented chain of reasoning, with the limits of the available evidence made explicit.

6.3. Detection and attribution

AI-assisted spectral analysis, multi-sensor fusion, and pattern recognition may improve the speed and sensitivity of chemical detection and identification. Their operational and evidentiary relevance, however, depends on validation against reference materials and representative field or laboratory conditions, including relevant instruments, environmental variables, chemical classes, and known sources of error. A model that performs well for a defined family of chemicals or a particular instrument configuration should not be presumed reliable beyond those boundaries. AI-assisted analysis may provide a time-sensitive field indication that informs immediate protective action, but confirmatory analysis through established sampling, quality-control, designated-laboratory, and reporting procedures remains essential.[34] The OPCW Laboratory and Designated Laboratories should have a central role in developing, testing, and evaluating such applications, with external scientific advice sought where needed.

Chemical-space mapping, retrosynthesis, and prediction of properties such as toxicity, stability, spectral behaviour, or detectability present a more anticipatory use case. These tools could strengthen preparedness by identifying chemicals or pathways of potential concern that are not yet central to routine practice. However, they may also influence threat perception, analytical targeting, and priorities for scientific work. They should consequently be restricted to defined purposes, validated with particular care, and accompanied by clear statements of assumptions and limitations. In alleged-use investigations under Article IX and Part XI of the Verification Annex, AI-supported steps must be independently reviewable and embedded in a documented factual chain. An AI tool may assist analysis, e.g. through retrosynthesis computations in an IAU by mapping possible synthetic pathways that explain the presence of certain markers that are useful for attribution purposes, but it cannot replace expert interpretation or convert probabilistic inference into attribution.

Human oversight must be substantive, which means that inspectors, analysts, and laboratory experts need interfaces that reveal uncertainty and limitations, training that enables them to question decision steps taken rather than defer to AI systems, and records showing the degree to which and how AI outputs did affect subsequent analysis. This is essential to the impartial execution of TS responsibilities as well as to States Parties' confidence that the organisation's functions are performed fairly. Transparency requires controlled explainability for authorised reviewers, auditable records, and enough procedural clarity to demonstrate that AI has not displaced treaty-based standards of evidence, confidentiality, or professional judgement.

7 Assistance and protection under the CWC

Article X provides a distinct operational setting for AI assurance because relevant tools may be used in training, preparedness, detection support, incident information management, triage support, and resource coordination under time pressure. The central domain-specific requirement is operational reliability in support of national emergency response systems: AI tools must be tested for the conditions in which responders and national authorities would actually use them, including incomplete information, variable connectivity, unfamiliar equipment, and rapidly changing circumstances.[35]

[34] Jerome Workman, "AI-Powered Near-Infrared Imaging Remotely Identifies Explosives," Spectroscopy Online, 11 June 2025.

[35] See OPCW, "Capacity Building: Assistance and Protection Programmes" and "Ensuring Preparedness."

7.1. A high-assurance response function

Any Article X tool should have a defined emergency-support function, named users, task-relevant performance evidence, and explicit limits on use. It must provide manual fall-back and ensure that exercises, deployments, and near misses feed into structured review. A recent review by the European Commission's Scientific Advice Mechanism finds that AI can significantly strengthen situational awareness, forecasting, and training in emergency and crisis management, but only where it is coupled with robust risk-assessment and AI literacy for responders.[36]

7.2. Detection, triage, and response

In an acute chemical emergency, the most useful AI support may not be direct chemical identification. Systems could assist clinicians and responders in recognising toxidromes and other patterns of exposure, combining reported symptoms, observable effects, environmental information, and available sensor data to prompt further questions, protective measures, or treatment pathways. Such support can be valuable when the responsible chemical has not yet been confirmed, but it must be treated as decision support rather than autonomous diagnosis or treatment allocation. It should clearly communicate confidence, uncertainty, and instances in which the observed case falls outside the system's validated experience or training data, so that users understand when escalation to specialist judgement is required.

AI-enabled detection, identification and analytical support offer clear potential benefits, but they must be closely controlled. Pattern-recognition tools may help interpret spectral information, combine multiple sensors, distinguish possible chemical hazards from background signals, or flag readings requiring expert attention. In a chemical emergency, this could reduce cognitive burden and accelerate protective decisions. Nonetheless, a false negative may expose responders or civilians to harm, while a false positive may misdirect scarce resources, create unnecessary alarm, or undermine confidence. Models must therefore be validated against representative agents, toxic industrial chemicals, environmental conditions, instruments, and operational scenarios. They should be used alongside established sampling, confirmatory analysis, medical, and incident-command procedures rather than as an autonomous source of identification. A systematic review of AI applications in disaster governance highlights tools for hazardous chemical identification, geographic information system-based planning platforms, and social-media triage systems, but emphasises that their effectiveness depends on governance structures, training, and data-management practices tailored to local conditions.[37]

The same restraint applies to AI-supported triage and resource allocation. Tools may help assemble information on casualties, available protective equipment, decontamination needs, transport routes, and the location of specialist assets. However, they should not autonomously determine who receives treatment or protection, nor should they embed unexamined assumptions that systematically disadvantage particular communities, regions, or linguistic groups. A systematic review of AI-based triage in emergencies and disasters finds that automated triage systems can reduce response times and optimise resource allocation, but also warns of biases and context-dependence that may systematically disadvantage particular communities if left unexamined.[38] In CWC Article X operations, fairness is inseparable from effectiveness: States Parties need confidence that assistance decisions rest on transparent, needs-based criteria, with human decision-makers accountable for any prioritisation. Systems should therefore be tested for

[36] Science Advice for Policy by European Academies (SAPEA), "Artificial intelligence in emergency and crisis management: Rapid evidence review report," Munich: SAPEA, 2025.

[37] Pirhossein Kolivand et al., "AI applications in disaster governance with health approach: A scoping review," *Archives of Public Health*, 83, 2025.

[38] Azadeh Tahernejad, Ali Sahebi, Ali Salehi Sahl Abadi and Mehdi Safari, "Application of artificial intelligence in triage in emergencies and disasters: a systematic review," *BMC Public Health*, 2024, 24: 3203.

differential performance across relevant contexts and accompanied by simple, usable explanations for authorised responders.

7.3 Training for meaningful human oversight

Training is a durable form of Article X assistance and a principal means of ensuring that AI remains under meaningful human control. The OPCW's existing training cycles, tabletop exercises, specialised courses, and workshops provide a foundation for this work; AI-supported simulation and virtual environments could also be used, where appropriate, to support preparation for Article VI industry inspections and other routine implementation activities. Any use of virtual models derived from real chemical-production facilities would, however, require careful treatment of confidentiality, export-control, commercial, and security concerns. AI-supported simulations and extended-reality environments could make exercises more realistic, adaptive, and accessible, including for scenarios involving chemical warfare agents and toxic industrial chemicals. They can help responders practise recognising uncertainty, coordinating across agencies, and deciding when automated advice warrants further confirmation. The aim should not be to train personnel to follow an AI system, but to train them to evaluate it as one source of information within a disciplined emergency-response process.[39]

Training systems must themselves be assured. Scenario design should be scientifically sound, culturally and linguistically accessible, and suitable for the infrastructure available to participants.[40] Offline or low-bandwidth options are important so that advanced simulation does not become a benefit confined to highly resourced States Parties. Courses should include AI literacy: data provenance, system limitations, bias, cybersecurity, confidential-information handling, incident reporting, and manual fall-back procedures. A systematic review of AI-based emergency response systems in smart infrastructure further underscores the importance of reliability, interoperability, and governance mechanisms to prevent over-reliance on automated decision support.[41] This aligns with Article X's requirement that States Parties facilitate the fullest possible exchange of information and means of protection, and with Article VII's requirement to adopt national implementation measures. It also advances the fairness principle according to which recipients of AI-enabled assistance must be able to scrutinise and govern the tools, not merely operate them. The above-mentioned SAPEA review also recommends integrating AI literacy into existing civil-protection training schemes, so that personnel learn to interpret automated outputs as one input among many and to recognise when further confirmation or escalation is necessary.[42]

7.4 Secure and equitable cooperation

Article X assistance will often require rapid exchange of sensitive operational, medical, and technical information. Secure architecture is therefore indispensable. Systems should employ role-based access controls, encrypted and auditable communication, data-minimisation rules, retention limits, and procedures for separating operational information from model training or external vendor environments. TS duties under Article VIII(5) to protect confidential information and to conduct verification in the least intrusive manner provide a relevant institutional benchmark, even when the immediate function is protective rather than fact-finding. Procurement arrangements should require providers to document data handling, security controls, model updates, incident cooperation, and the Organisation's ability to suspend a tool where assurance conditions are no longer satisfied.

39 OPCW, "Training Cycles"; OPCW, "Article X Workshops."

40 Kumpol Saengtabtim et al., "Harnessing generative AI for enhanced disaster management: a systematic review," *Big Earth Data*, 2025, 9(4), pp.651-675.

41 Ammar Bajwa, "AI-based emergency response systems: A systematic literature review on smart infrastructure safety," *American Journal of Advanced Technology and Engineering Solutions*, 1(1), pp.20-40.

42 SAPEA, "Artificial intelligence in emergency and crisis management," p.76.

Finally, Article X should be implemented in conjunction with Article XI's commitment to non-discriminatory scientific and technological cooperation. AI may improve collective preparedness only if the necessary infrastructure, training, documentation, and support reach States Parties with differing levels of resources and technical maturity. Shared exercises, peer-to-peer partnerships, multilingual guidance, jointly developed validation protocols, and regional support networks can reduce the risk that AI creates a new protection divide. The ChemTech Centre and the International Cooperation and Assistance Division can provide institutional platforms for this work.

8 Capacity building, fairness, and the technology divide

Access to AI does not itself create an operational capability in the physical world. Meaningful capacity depends on the infrastructure, equipment, connectivity, maintenance, reference materials, qualified personnel, regulatory permissions, and supply arrangements needed to act on AI-supported information. OPCW ICA programmes should therefore address governance capacity, technical infrastructure, and sustainable access alongside software or training. Priority areas include data stewardship, AI misuse awareness, procurement and vendor assessment, validation literacy, secure deployment practices, the ability to recognize limitations, and contingency planning for changing or obsolete systems. Twinning and regional peer-learning arrangements can connect National Authorities, laboratories, academic institutions, and relevant technical communities. In this context, a recent policy analysis emphasises that a 'Global South pivot' in AI governance should prioritise digital inclusion, South-South cooperation, and investments in local capabilities, so that equity concerns are addressed structurally rather than through ad hoc technical assistance.[43]

8.1 From access to governance capacity

The SAB TWG on Artificial Intelligence offers a practical route for addressing the technology divide. Its recommendations envisage a dedicated programme for States Parties covering emerging AI capabilities, recognition of malicious and accidental misuse, data governance and management, and AI-supported simulation and communication tools for chemical safety and emergency response. Properly understood, this is not a peripheral assistance package. It is an institutional safeguard against unequal preparedness in a rapidly changing technical environment. The OPCW's existing capacity-building portfolio, including laboratory assistance, laboratory twinning, fellowships, courses, and support through the Centre for Chemistry and Technology (ChemTech Centre), provides an established platform through which such assistance can be organised. Work on verification methods for prospective international AI agreements explicitly cites organisations such as the OPCW and IAEA as models for combining training, inspection powers, and technical expertise, suggesting that future AI regimes will likely rely on capacity-building mandates similar to those of existing arms-control bodies.[44]

The key point is that capacity building must go beyond technology transfer. Responsible AI use depends on data curation, provenance, stewardship, ownership, reliability, protection, access controls, documentation, and the ability to test a system's claims against local conditions. A recent proposal for eliminating the AI digital divide argues that local technical, operational, and educational capacity must be built before complex compliance duties are imposed.[45] Resource-constrained institutions that receive software without these supporting capabilities may inherit new vulnerabilities: unreliable outputs, inappropriate reuse of sensitive data, vendor dependence, or systems that they cannot validate, adapt, or maintain. Project design should consequently begin with a context assessment covering infrastructure, connectivity, local expertise, language

43 Deepak Maheshwari, *AI's Global South Pivot: Equity, Ethics and Ecology*, Policy Brief No. 225, Waterloo: Center for Governance Innovation, February 2026.

44 Akash R. Wasil et al., "Governing dual-use technologies: Case studies of international security agreements & lessons for AI governance", arXiv preprint, September 2024.

45 Freya Gulamali et al., „Eliminating the AI digital divide by building local capacity," *PLOS Digital Health*, 4(10), 2025.

requirements, available reference data, maintenance capacity, and national legal constraints. A tool that is technically sophisticated but cannot be maintained, updated, independently assessed, or integrated with the available physical and institutional infrastructure does not constitute meaningful capacity building. Sustainability plans should address likely model and software change, compatibility with existing systems, maintenance responsibilities, update costs, user support, and the continued availability of non-AI or alternative arrangements if a tool becomes obsolete or unsuitable.

8.2. Laboratories and shared validation

AI-enabled laboratory assistance is a particularly important case. The Laboratory Assistance Programme seeks to strengthen national capacities for chemical analysis and monitoring, while the Laboratory Twinning Initiative connects aspiring laboratories with experienced partners, including laboratories that have held OPCW designation. AI tools for spectral interpretation, multi-sensor fusion, and predictive analysis could help laboratories develop sophisticated capabilities more rapidly, especially if combined with access to ChemTech Centre training and expert networks. Yet, a method should not be disseminated merely because it performs well in a well-resourced development environment. Its reliability must be assessed across the instruments, sample types, technical conditions, and staff capacities that partner laboratories actually possess.

Cooperation projects should use shared validation and benchmarking arrangements, but should avoid assuming that one detailed operating procedure can fit every laboratory. The relevant model is closer to performance-based recommended operating procedures: common requirements define what an AI-assisted method must demonstrate, while each laboratory develops institution-specific procedures compatible with its instruments, infrastructure, sample types, and staff competence. Designated Laboratories, assisting laboratories, and recipient laboratories could assess tools jointly using appropriately authorised data. This approach adapts established laboratory-quality principles to AI without implying that automated outputs can replace scientifically accountable analysis. It also spreads the capacity to question and audit tools, rather than centralising that capacity in a small technical elite.

8.3 Training and preparedness

AI-supported training offers another way to broaden rather than narrow capacity. The ChemTech Centre's training and technology functions, together with regional courses and assistance-and-protection programmes, could use extended-reality simulations, language tools, and knowledge-management systems to provide realistic exercises in chemical safety, emergency response, and national implementation. They could also support preparation for Article VI industry inspections, including scenario-based training on facility-specific context, inspection questions, and the limits of AI-supported information. Such uses should be developed with close attention to confidentiality and to the fact that inspections, facilities, and national implementation arrangements are not identical. Such applications are closely connected to Article X, which includes advice, detection, protective measures, decontamination, and treatment within the meaning of assistance. They can also support Article VII, under which States Parties must adopt national implementation measures and designate a National Authority as a focal point for liaison with the OPCW.

However, accessibility cannot be assumed. Training systems should be available in multiple languages, designed for low-bandwidth or offline use where necessary, and tested for pedagogical effectiveness across varied institutional settings. Participants should receive AI literacy as well as task training: how to recognise erroneous or manipulated outputs, understand model limitations, protect sensitive data, report incidents, and decide when an automated recommendation must be overridden. The UNESCO Recommendation on the Ethics of AI similarly stresses inclusion, fairness, transparency, human oversight, and AI literacy.[46] For the OPCW, these general principles should be

[46] UNESCO, Recommendation on the Ethics of Artificial Intelligence, 2021.

operationalised through the Convention's specific commitments to assistance, peaceful cooperation, and non-discriminatory technological development. The Advisory Board on Education and Outreach could contribute within its existing mandate by advising on AI literacy, accessible communication, and the socialisation of responsible-assurance practices across relevant OPCW training and outreach activities.

8.4 Twinning and co-design

The TWG's proposed twinning initiative addresses the need for durable peer learning. Building on the existing Laboratory Twinning Initiative, partnerships should link aspiring laboratories with experienced Designated Laboratories where the purpose is to strengthen the capacities relevant to designation; some, though not all, Designated Laboratories may also be well placed to share relevant experience with AI-assisted analytical methods. The existing laboratory-twinning model offers a credible institutional foundation for this approach.[47]

Twinning should therefore be organised as co-design rather than one-way transfer. Recipient laboratories, National Authorities, emergency responders, and regional experts should participate in defining priorities, assurance criteria, documentation, and success measures. Regional workshops can identify trade-offs between analytical performance and infrastructure constraints, and can ensure that AI-related assistance is compatible with both OPCW confidentiality requirements and domestic legal frameworks. Accessible documentation, continuing technical support, and channels for user feedback are essential. They enable participants to report failures, seek clarification, and contribute evidence that improves the tool and its governance over time.

8.5. Fairness in implementation

Capacity building is therefore integral to regime legitimacy. AI-enabled methods for analysis, preparedness, or threat assessment should not concentrate interpretive authority in a small group of technically advanced actors. Shared validation, multilingual documentation, AI literacy, peer learning, and sustainable local infrastructure can instead make AI a collective implementation capability. In this context, the SAPEA review of AI in emergency and crisis management recommends that capacity-building programmes include explicit modules on bias, over-trust, system limitations, and incident reporting, so that recipients can scrutinise and govern AI-enabled practices rather than merely operate them.[48] Such an approach can also help States Parties recognise both malicious and accidental misuse, reducing the risk that technological asymmetry becomes an asymmetry of security.

9 Conclusions

AI is a crosscutting condition of contemporary CWC implementation and of the OPCW's routine organisational practice. This paper has argued that the appropriate response is a risk-calibrated AI assurance architecture, understood as a lifecycle process for determining whether an AI-enabled system is trustworthy for a defined purpose and operating context. In this sense, AI assurance is a permanent governance requirement through which the OPCW can ensure that AI augments, rather than displaces, treaty-based expertise and accountability.

The practical standard for the introduction and use of AI systems should be structured restraint. AI applications bearing on verification, compliance-relevant analysis, chemical identification, or operational assistance must be treated as high-assurance systems whose use should be governed by the enhanced assurance requirements set out above.[49] For lower-risk administrative or

[47] See OPCW, "Capacity Building. Laboratory Twinning Initiative"

[48] SAPEA, "Artificial intelligence in emergency and crisis management," p.76.

[49] See section 3 above.

knowledge-management tools, proportional safeguards may suffice, but confidentiality and secure-use requirements should remain non-negotiable for AI utilization under the CWC.

An OPCW AI assurance framework should combine hierarchical and relational assurance. Hierarchical assurance procedures and tools make it possible to demonstrate that an AI system meets pre-defined requirements for legality, robustness, confidentiality, fairness, explainability, and accountability. They are especially necessary where AI contributes to a sensitive evidentiary or operational workflow. Relational assurance is equally important as it concerns the engagement, explanation, contestability, and feedback through which confidence is sustained among those who use or are affected by AI. It requires co-design with operational users, practical AI literacy, channels for reporting errors and questioning outputs, and managed transparency that explains safeguards and limitations without disclosing confidential data or sensitive methods. Formal compliance alone will not produce trust if relevant stakeholders cannot understand how AI is used, raise concerns, or see those concerns shape practice.

Implementation should be sequenced rather than simultaneous. In the *short term*, the Secretariat should establish an inventory of existing and proposed AI uses across all divisions, distinguishing ordinary productivity tools from systems that materially support verification, laboratories, assistance and protection, or international cooperation. Each system should receive an initial risk screen based on treaty function, data sensitivity, potential consequences, degree of automation, external dependency, operational-continuity requirements, and likely effects on States Parties. Interim guidance should establish immediate guardrails, including no unauthorised use of public AI services for confidential or operationally sensitive information; human responsibility for consequential outputs; documented approved uses; secure data handling; and procedures for incident escalation, and system withdrawal.

The OPCW should then conduct a small number of well-designed pilots. Their boundaries should be defined by operational usefulness with each pilot having a clearly specified purpose, representative authorised data, independent testing by people not involved in development, predefined success and failure criteria, user feedback, and a continuity plan. States Parties and relevant operational users should, where appropriate, participate in evaluation and provide real-world feedback. Suitable initial applications could include protected knowledge retrieval, AI-assisted translation in a secure environment, simulation-based training, or retrospective analysis of whether an AI system identifies declaration issues comparable to those previously raised by human reviewers. Verification-related pilots must not influence compliance-related decisions unless explicit assurance thresholds and the applicable legal and policy requirements have been met.

In the *medium term*, the Director-General should adopt an OPCW AI Assurance and Data Governance Policy. It should define CWC-alignment principles, roles and responsibilities, risk categories, baseline and enhanced assurance requirements, security and confidentiality controls, human-oversight expectations, documentation, audit, and incident-reporting obligations, and rules governing procurement and third-party systems. Practical quality assurance templates should support the policy: an AI Assurance Plan, impact and risk assessment, test report, model or system record, change-control form, and incident log. TS-wide project approval, procurement, information-security, and internal-audit processes should be updated so that AI assurance is built into design and contracting rather than added after deployment.

Domain-specific standard operating procedures should follow or be updated for an AI-enabled environment. In verification, they should place a premium on data handling, validation, advisory status, escalation, and expert review for declarations processing, anomaly detection, data fusion, mission support, and chemical identification. In assistance and protection, they should address reliability in field conditions, uncertainty displays, manual fall-back, instructor control over simulations and misuse-sensitive content. In international cooperation, they should require context-appropriate testing, multilingual and accessible documentation, recipient participation, sustainable support, and shared validation where AI tools are transferred or co-developed. Staff training – particularly for inspectors, laboratory personnel, programme managers, procurement staff, information-security specialists, and responsible system owners – is essential to transform formal arrangements into operational practice. A small AI Assurance and Data Governance Unit or cross-functional coordination mechanism, built on existing Secretariat capacity, could coordinate the work while leaving operational ownership with the relevant divisions. The SAB should continue to advise on emerging scientific and technological developments and novel uncertainties within its mandate; the ABEO could help develop AI-literacy, outreach, and confidence-building elements; and specialised external or designated-laboratory expertise should be sought where the task requires it.

In the *longer term*, AI assurance should be incorporated into routine OPCW governance and into structured engagement with States Parties. The Director-General could provide periodic high-level reporting to the Executive Council and Conference of the States Parties on the categories of AI use, safeguards, significant incidents, lessons from pilots, and capacity-building progress, while respecting the Convention's need-to-know confidentiality regime. The Executive Council's standing agenda item on emerging technologies could provide an existing forum through which a State Party champion, the Secretariat, or relevant advisory bodies bring forward AI-assurance issues for discussion and socialisation. This would allow shared principles and practical expectations to develop through established multilateral processes rather than presuming agreement on detailed operational specifications from the outset.

In sum, the OPCW should approach AI not as an innovation agenda detached from treaty implementation, but as a governance agenda in service of the CWC's object and purpose. A risk-calibrated AI assurance architecture, built around legal alignment, confidentiality-compatible design, validated performance, substantive human authority, accountable organizational ownership, and inclusive cooperation, can enable the OPCW to realize useful AI applications without sacrificing trust.

The CBW network for the comprehensive strengthening of norms against chemical and biological weapons (CBWNet)

The research project CBWNet is carried out jointly by the Berlin office of the Institute for Peace Research and Security Policy at the University of Hamburg (IFSH), the Chair for Public Law and International Law at the University of Gießen, the Peace Research Institute Frankfurt (PRIF) and the Carl Friedrich Weizsäcker-Centre for Science and Peace Research (ZNF) at the University of Hamburg. The joint project aims to identify options to comprehensively strengthen the norms against chemical and biological weapons (CBW).

These norms have increasingly been challenged in recent years, *inter alia* by the repeated use of chemical weapons in Syria. The project scrutinizes the forms and consequences of norm contestations within the CBW prohibition regimes from an interdisciplinary perspective. This includes a comprehensive analysis of the normative order of the regimes as well as an investigation of the possible consequences which technological developments, international security dynamics or terrorist threats might yield for the CBW prohibition regimes. Wherever research results point to challenges for or a weakening of CBW norms, the project partners will develop options and proposals to uphold or strengthen these norms and to enhance their resilience.

The joint research project is being funded by the Federal Ministry of Research, Technology and Space for six years (April 2022 until March 2028).

Author Information

Dr. Alexander Kelle is a Senior Researcher at IFSH, and at the Berlin Office heads the CBWNet project.

Acknowledgements

The author would like to thank Dr. Ralf Trapp and Dr. Jonathan Forman, who both served at different times as Secretary to the SAB. Both have provided, in their personal capacity, very useful comments on an earlier version of this paper. As always, any remaining errors are the author's responsibility.

Contact

Responsible for Content §55, Abs.2 RStV:
Dr. Alexander Kelle, IFSH
Reinhardtstraße 7, 10117 Berlin
E-Mail: CBWNet@ifsh.de

www.cbwnet.org
twitter.com/CBWNet

ISSN (Online): 2751-4501

With funding from the:

Federal Ministry of Research, Technology and Space