Confidence and Trust in Nuclear Disarmament Verification

December 2025

Introduction

The concepts of trust and confidence feature prominently in discussions about international relations generally and nuclear disarmament verification specifically. In Phase III, the IPNDV addressed this topic, distinguishing between confidence and trust to help foster increased nuance and precision. This effort explored how to measure confidence in the different elements of the verification system, as well as in the verification regime itself. Ideally, over time, the effective implementation of a verification system will build increased confidence as well as trust.

To help the parties to arms control agreements make informed decisions and judge whether the other parties can be relied upon, these agreements typically include various verification, compliance resolution, and decision-making mechanisms. Data requirements and the processes, procedures, techniques, and technologies (PPTT) to collect data form the *verification system.*

Definitions

Confidence in verification of nuclear disarmament is understood as being founded on an evidence-based assessment on the part of the inspecting entity as the basis for evaluating compliance with the obligations of an agreement. *Trust,* by contrast, is understood as a personal, subjective assessment made by individuals.

Verification

Verification encompasses the technical elements of monitoring and inspection as well as information processing and evaluating compliance. A verification system should be designed to address the specific objectives of an agreement. It is designed to establish the degree of compliance with the specific terms of an agreement. Verification will likely take place over the lifetime of the agreement and could include a multilateral inspecting entity.

The aim of verification is to increase confidence that an agreement is being fully implemented by parties to it, with the opportunity to convincingly demonstrate their compliance and to detect non-compliance, thereby deterring cheating.

Confidence Assessment

The design of the verification system needs to provide a sufficient level of confidence in all of its parts—in every piece of equipment, measurement, inspection, as well as data processing and storage. At the same time, building confidence should be seen as a continuing, layered process.

An assessment of confidence produced by an inspecting entity should be robust and repeatable. Confidence assessments are objective in the sense that they are reducible to the application of standardized PPTTs and rules for data interpretation.

Subjectivity

Although the ideal confidence assessment is completely objective, subjective assessments and biases could affect both the design and implementation of a verification system.

Design

When developing procedures for using a technology, for example, a technologist has certain perceptions of their technology's capabilities and performances that may not be repeatable in all environments. This could, for example, result in overestimating the reliability of their equipment. Additionally, in evaluation subjectivity exists in the weighting scheme based upon the evaluator's perception of the environment (see paragraph on "Assessing Confidence Using Influencing Criteria" below). Lastly, there is subjectivity in the determination of what output is necessary to deliver confidence.

Designing a verification system will have to take several real-life practicalities into consideration:

- Resource constraints (e.g., the numbers of inspectors, costs of equipment, etc.)
- Time constraints (e.g., the amount of time that is agreed in the agreement for the inspectors' access to the sites)
- Access constraints (e.g., host safety or security restrictions that limit access or restrict inspection activities in specific environments)

Hence, verification involves *compromises* (e.g., on numbers and locations of inspections) and *trade-offs* between data gathering methods (e.g., between on-site inspection and continuous monitoring). The trade-offs and compromises will manifest themselves in the negotiated protocols of a specific agreement, as well as throughout the implementation of the treaty.

Implementation

An inspecting entity may also experience unexpected disruptions of monitoring or inspections. For example, weather, accidents, and other events may interrupt planned verification activities or cause ambiguities in technical measurements and inspection results.

Given these factors and the potential fallibility of verification mechanisms, the aim is to assess what is necessary to achieve a *sufficient level* of confidence. Persons involved in any verification process inevitably must conduct their work in a space that falls short of absolute certainty. Outcomes represent expressions of confidence in the degree to which it is believed that the inspected party is adhering to agreed rules. There is not one set level of confidence. What constitutes a *sufficient level* of confidence will vary, based on the factors described above.

However, there are numerous different mechanisms that could be used to measure elements of a verification regime and provide quantifiable data on the effectiveness of those elements in a single application, even though quantification of overall confidence may still be elusive.

Bolstering Confidence

To provide evidence that can be used by assessors to reinforce their perception of confidence, a variety of tools are available.

Random Selection Methods

To create an effective approach for evaluation of confidence over time, random selection poses a reasonable option to deal with the limitations described above. Random selection is a strong measure by itself, but there are mechanisms that can augment random selection by identifying tools and applications that, when applied to random selection, can create an effective approach for evaluating confidence. Inspections that involve random selection measures can serve to develop a growing body of evidence which, over time, increases the estimation of confidence. The ability to observe consistency in behaviors, processes, documentation, and results lends more credence to the host's demonstrations of compliance through openness and transparency.

Radiation Measurements

Radiation measurements enable inspectors to build confidence through the detection of presence or absence of special nuclear materials (SNM) used in nuclear weapons. For the hosts, accepting measurements risks revealing sensitive information. Therefore, the part of an agreement that deals with radiation measurements must developed to adequately address the concerns of all parties. This calls for cooperation between technical and non-technical experts. The measurements may require using information barriers to protect sensitive information.

Radiation measurements are not conducted in isolation. Instead, they are paired with tamper-indicating seals, tags, and other chain-of-custody tools; visual inspections; data analysis (includes cross-comparisons of data); reporting; and secure long-term data storage to form a verification system that will be able to adapt to changing situations throughout the life of an agreement. Note that optimal instruments for absence and presence measurements may be different.

Measurement-based confidence depends on the amount, quality, and processing of collected data; the equipment used; and measurement geometries. It may be limited by the nature and presentation of the source (treaty accountable item (TAI) and its container) that may not be precisely known. Please see Annex A for a more detailed discussion on measurements.

Systems Approach

An additional measure to bolster confidence is to apply a *systems approach* to verification. Such an approach considers a state's nuclear weapons-related infrastructure and related technical capabilities as a whole, allowing analysts to investigate whether the state's nuclear weapons enterprise (NWE) operates consistently with the requirements of an applicable agreement.

One reason for looking at the NWE as a whole is the realization that inspection resources are limited, and it is impossible to verify all things at all times. It may not be possible to verify all individual movements of items within the enterprise. By identifying verifiable sub-systems and understanding their relationships, it should be possible to see behavior consistent with what has been declared across the system as a whole and build confidence in a state's compliance with its treaty obligations. The systems approach to nuclear disarmament verification explores how various sub-systems of the NWE should operate under treaty requirements.

The NWE would include all TAIs, the facilities and supporting infrastructure that supports them, and operations and processes involving accountable items and facilities.

Although the IPNDV 14-step model has been primarily concerned with the potential for clandestine *diversion* of declared nuclear warheads and their associated components as they move through to dismantlement and disposition, the broader systems approach considers a states' ability to *acquire* undeclared TAIs, to quickly *break out* from the regime, or to *reconstitute* militarily-significant capabilities.

It should be noted that there is a distinction between how the NWE will be seen operating as a system by the host and how the system, and sub-systems, will be seen for verification purposes that are based on treaty definitions. The systems approach is a tool for achieving *high-level verification objectives* and can be used to develop *implementation-specific verification objectives*.

A statistical approach can then be used to distribute verification resources among implementation-specific verification objectives.

The Human Factor

Within the scope of the information that an inspecting entity would be allowed to gather, the level of confidence generated should, in theory, be directly related to the technical strength of the evidence acquired and the extent to which it stands up to scientific scrutiny. The elements of subjectivity in verification system design and implementation mean, however, that outcomes have the potential to be influenced by psychological factors.

Despite a lack of abundant quantitative evidence, having human beings involved in decision making related to nuclear disarmament verification has a clear advantage: humans have the ability to perceive non-quantifiable behaviors and take in-the-moment context into account. Hypothetically, a human inspector might be able to discern changing atmospherics in the relationship between inspectors and hosts, possibly indicating that the inspection is being conducted in a different manner, which, in turn, could have implications for evaluations of consistency. Simple variations in how two different individuals may perform activities might not be noticed technically but might be noticed by experienced inspectors who had been at the site

for previous inspections. In theory, humans can fill the gap between technical verification and the perception of confidence by providing that valuable context. Humans can assess consistency through visual or other cues and determine their relevance without relying solely on the presence or lack of technical information, which may be restricted due to proliferation concerns.

Despite these benefits, it is also important to understand the potential disadvantages of the human factor. These include the influence of bias, that is, a *prejudice in favour of or against one thing or group compared with another, usually in a way considered to be unfair.* It is important that we know how to recognize it, how to mitigate it, and why it is an essential element of the human factor when assessing compliance with disarmament measures. Although evidence of the role of bias in disarmament verification is primarily anecdotal and limited due to the often-classified nature of inspections, acknowledging even the potential for bias to influence assessments of compliance is vital in developing a verification regime.

Conscious bias refers to prejudice that is known and acknowledged by relevant individuals. *Unconscious bias* describes the subliminal or unaddressed beliefs, values, and opinions that affect how we interact with the world and one another.

Common biases include prejudice anchored in ideas about race, gender, ethnicity, economic class, sexual orientation, educational background, and age. It is also important to recognize that individual biases are developed in connection with wider norms and discourses such as systems of patriarchy and colonialism. Bias, whether conscious or unconscious, may contribute to the general sense that the inspected state is hiding important information or is intending to deceive. Please see Annex B for further discussions on biases in nuclear disarmament verification.

The challenge is to recognize that biases exist, and design methods that allow for as much of an impartial assessment of compliance as possible—not by disregarding the "human factor," but by understanding how it can be successfully used and mitigated within verification systems.

A measure that would limit the negative influence of the human factor is using indicators to measure confidence. This entails operationalizing the "sufficient level of confidence" necessary for each factor going into to a verification system, as well as for the system as a whole.

Measuring Confidence

IPNDV has examined several methods to investigate how to measure confidence. This chapter describes the initiatives explored in Phase III.

Concepts of Operations

Although an important factor in verification, technology does not provide the complete answer. In its development of Concepts of Operation (CONOPS), the IPNDV introduces indicators to measure confidence in a mix of PPTT, assigning a "confidence value" of low, medium, or high. Further work is needed on developing the criteria for each level.[1]

[1] CONOPS are developed for specific verification activities and ask the user to describe the objective of an activity; the needed declarations notifications; and data required, activities to be carried out, required equipment, and any technology requirements.

A Pathway for Equipment Confidence Through Modifications of Concern

The IPNDV was also introduced to a methodology aiming to quantify confidence in equipment. The aim of this methodology[2] is to develop and demonstrate a repeatable approach to more objectively quantify confidence in evaluated equipment proposed for use in future verification regimes. Input to this process is proposed to consist of the equipment and use case parameters within a verification regime. Output from the process is a prioritized list of methods to defeat inspection equipment and mitigation strategies ranked across a series of metrics. This output is available to stakeholders to inform which attack mitigations have the highest priority to implement, and the remainder not selected represents residual risk to the equipment that will be used to quantify confidence in the equipment. Additional output of the process is information to guide future joint inspection activities based on lessons learned, potential attacks, and required capabilities on inspection tools to mitigate selected attacks defined earlier in the process. The process involves a series of steps:

1. Identify subject matter experts (SMEs) to support remaining steps
2. Identify modifications of concern (MoC)
3. Identify hypothetical attacks that exploit MoCs
4. Evaluate, score, and prioritize hypothetical attacks against a defined set of metrics
5. Identify mitigation strategies and inspection techniques
6. Identify security critical components (SCCs)
7. Rank attack/mitigation pairs based on score
8. Determine Equipment Confidence One important note about this methodology

It is not intended to replace a full vulnerability assessment (VA) or a hands-on evaluation of equipment using national capabilities (e.g., 30-day evaluation). This approach is based on a tabletop approach using subject matter expertise to conceptualize attacks and associated solutions. The choice of SMEs can help to identify the fidelity or amount of confidence that may be achieved by going through this methodology. It can enhance or augment these other activities and focus scarce resources where they make the greatest impact.

As an example, a wide range of activities could be performed on a piece of equipment as part of a 30-day examination. By following this proposed methodology, these activities will focus on an equipment examination to prioritize potential adversary pathways to defeat the equipment, identify what evidence the conceptualized pathways may leave behind, and therefore what examination techniques to apply to determine the presence or absence of that evidence. For the 30-day evaluation, this gives insight into the security of the equipment that can be provided to stakeholders as a recommendation by SMEs. For in-field inspection activities, output from the methodology will highlight critical inspection activities, areas to inspect, and an awareness of

[2] This Confidence Methodology is under development by the joint U.S.-U.K. Authentication Certification Working Group (ACWG).

what to look for. The presence or absence of evidence at this point creates confidence in the equipment to be used during an on-site inspection.[3]

Bayesian Analysis: Quad Charts

Whereas confidence is commonly assessed quantitatively using the results of statistical data as mentioned above, it also has a qualitative element that has, until, now gone largely unrecognized. The IPNDV developed a diagramming tool that uses an eight-vector Bayesian Network Analysis process to qualitatively evaluate individual PPTT elements as to the degree of effect each has against the criteria being validated. This criterion could be for example, "confidence in the resilience of the monitoring and verification regime against diversion," or "confidence in the robustness of the regime against system failure."

More specifically, for this assessment, Bayesian Network Analysis is used to evaluate each PPTT element on a scale of influence (0, .3 (low), .6 (moderate), .9 (significant)), in relation to its individual impact in successfully achieving the validation criterion. As part of that analysis, the IPNDV developed a set of overarching influencing criteria (IC) that focus on the cumulative relationship between the inspecting entity and host resulting from prior historical interactions. These IC define the verification context and show how confidence can change over time and in response to relational changes. Within the overall Bayesian Network Analysis, these IC also are scored and then weighted, providing an average of overall relational confidence. This overall relational confidence is then reviewed as to its effect on the PPTT quadrants and regime. Finally, as part of the analytic process, the PPTT element scores are then rolled up by quadrant and averaged and then combined and re-averaged with the resulting IC score, providing a total quadrant weighted value. This provides a tool by which inspectors can identify the importance of individual and groups of PPTT during the lifetime of an agreement, and the effect of confidence during that period in response to unexpected changes either in IC scores or PPTT.

Examination of this methodology through a series of mini exercises, identified that technical confidence may be limited at the start, but will likely increase over time and with repetitive result consistency. It was also noted that the effect of stable, enduring ICs could be a mitigating factor in situations where technical problems arise. This mitigation was less significant early in the mini exercise scenario when participants were asked to consider this to be their first verification visit, but over time the stability of the IC became more impactful in bolstering confidence, even in the face of unusual or unexpected technical failures.

Trust

If we understand *confidence* as an assessment based on standardized rules and procedures, *trust* can be understood as a subjective, personal perception of the extent to which someone or something can be relied upon. In contrast to confidence assessments, trust assessments are private.

[3] Jacob Benz, Neil Evans, Neil Grant, Jon Warner, Tom Weber, and Joseph Froeschle, *A Pathway for Equipment Confidence Through Modifications of Concern* (2023).

Defined as a private perception, trust can initially be vague and difficult to influence or operationalize in disarmament verification. Yet trust is critical to any successful nuclear disarmament process. After all, policy choices are made by political leaders who each hold their own private views and opinions. These views will of course be informed by a range of external factors—for example confidence assessments produced by verification bodies—but may be ultimately and inescapably personal to the individuals in question. This may be true, for example, for hypothetical decisions about whether to join or withdraw from arms control agreements, to pursue options for cheating in an inspection regime, or to challenge other parties to an agreement. It has been convincingly established that the arms control and disarmament agreements initiated by U.S. President Ronald Reagan and Soviet Premier Mikhail Gorbachev during the second half of the 1980s came about in large measure as a result of the development of interpersonal trust between the two leaders.[4]

In general, the assumption should be that the production by a verification body of assessments indicating high confidence in compliance *should* foster trust in that compliance among relevant stakeholders. There are, however, exceptions to this rule. For example, during 2002–2003, a number of influential policymakers remained insistent that Saddam Hussein's Iraq had weapons of mass destruction (WMD) despite the production by the United Nations and International Atomic Energy Agency (IAEA) of a series of reports indicating increasing levels of confidence that Iraq was not in possession either of usable WMD or significant WMD production capabilities.[5] In fact, the increasing evidence of absence was interpreted by some as "simply a sign that he [Hussein] had gotten even better at hiding them [the WMD] from us."[6]

Determining Compliance

The design and implementation of a nuclear disarmament verification regime "requires that a principal locus of verification authority be agreed by the parties."[7] This premise begs at least two questions. First, who has the authority to carry out inspections or other measures to collect evidence about states parties' compliance with the obligations of an agreement? And second, who has the authority to conclusively decide whether a particular party is or has been in compliance with said rules? The latter may amount, for example, to determining whether a given number of technical anomalies such as broken seals or inconsistencies in reporting amount to non-compliance in the broader sense, that is, in the sense of warranting high-level attention

[4] See, e.g., Nicholas J. Wheeler, *Trusting Enemies* (Oxford: Oxford University Press, 2018).

[5] Robert E. Kelley, "Twenty Years Ago in Iraq, Ignoring the Expert Weapons Inspectors Proved to Be a Fatal Mistake," Stockholm International Peace Research Institute, March 9, 2023, https://www.sipri.org/commentary/essay/2023/twenty-years-ago-iraq-ignoring-expert-weapons-inspectors-proved-be-fatal-mistake. Inspections were carried out by the IAEA and United Nations Monitoring, Verification and Inspection Commission (UNMOVIC).

[6] Kenneth M. Pollack, cited in Elizabeth Shelburne, "Weapons of Misperception," *The Atlantic*, January 2004, https://www.theatlantic.com/magazine/archive/2004/01/weapons-of-misperception/303110/.

[7] IPNDV, "Verification of Nuclear Disarmament: Insights from a Decade of the International Partnership for Nuclear Disarmament Verification," June 2024, p. 15, https://www.ipndv.org/wp-content/uploads/2024/06/IPNDV-Capstone_FINAL-1.pdf.

and/or the triggering of treaty-mandated enforcement mechanisms. Existing and historical international agreements offer a range of possible answers to these questions.

With respect to the collection of evidence about states' compliance, arms control and disarmament agreements typically rely on one or more of the following stylized measures:

- The application by individual states, on a discretionary basis, of national technical means of verification
- Treaty-mandated verification measures carried out by individual states (such as the monitoring and inspection activities mandated by New START)
- Verification activities carried out, coordinated, and conveyed by an international organization (such as the IAEA, Organisation for the Prohibition of Chemical Weapons (OPCW), or Comprehensive Nuclear-Test-Ban Treaty Organization)

With respect to drawing final conclusions about whether specific states are or have been in compliance with treaty obligations, arms control and disarmament agreements invariably place the locus of authority with one or more of the following actors:

- Individual states parties (as in the case of bilateral U.S.-Soviet/Russian arms control)
- One or more intergovernmental bodies set up under an international organization (as in the case of the OPCW) and/or
- The staff or technical secretariat of an international organization (as in the case of the IAEA)[8]

Many international agreements, including arms control and disarmament instruments, such as New START and the Biological Weapons Convention, also provide for official consultation arrangements that allow states to raise or work through compliance concerns in a cooperative fashion. For example, the Joint Compliance and Inspection Commission under START allowed the treaty parties to discuss anomalies detected during recurring monitoring and inspection activities.

The allocation of verification authority can have significant consequences for the overall verification enterprise. For example, institutional design choices could have major implications for the ability of the entity charged with making final decisions about compliance to use or receive information from national intelligence agencies. Institutional dynamics could also condition the verification regime's ability to follow up on cases of concern, for instance through extraordinary verification missions. Finally, the style and manner of reporting on verification findings would likely be dictated, in large measure, by the institutional makeup and technical proficiency of the entity charged with making decisions about compliance.

[8] In the case of the IAEA, non-compliance has been established six times. On four occasions, the IAEA Board of Governors made the call. On two occasions, the IAEA Secretariat made the call. See Olli Heinonen, *IAEA Mechanisms to Ensure Compliance with NPT Safeguards* (Geneva: UNIDIR, 2020). The IAEA Statute tasks the IAEA's "staff of inspectors" with "determining whether there is compliance." IAEA Statute, Article XII(C).

It is often argued that verification regimes should be anchored in norms of efficiency, scientific integrity, and impartiality. Many would maintain that this implies vesting significant verification authority in a capable technical organ unencumbered by political interests and allegiances beyond the fulfilment of its mandate. In this view, a technical verification body would be less likely than sovereign states to be influenced by concerns with trade, the balance of power, or alliance politics; a technical verification agency would be free simply to follow the evidence. That being said, history suggests that involvement in politically salient decision making—such as the determination of states' compliance with treaty obligations—can open even technical agencies to charges of bias or partisanship. In turn, such charges or perceptions of political bias could conceivably serve to constrain inspector access to aggrieved states parties, undermining the verification enterprise. One could also make the argument that decisions about compliance—or certainly any execution of enforcement mechanisms—should not be made on the basis of objective criteria but rather a contextual political analysis of the likely consequences of alternative courses of action. The need for supranational technical authority may also vary depending on the breadth and depth of the disarmament treaty under negotiation. Negotiators must take these and other considerations into account when designing any future nuclear disarmament verification regime.

Conclusion

A key purpose of producing confidence assessments about states' compliance with arms control or disarmament agreements is to inform the trust assessments of relevant partners.

As suggested above, for any verification system to retain credibility, relevant stakeholders must have a sense that the techniques, equipment, and data used to produce confidence assessments are sound and reliable. To the extent that the inspecting entity maintains and applies standardized rules and procedures to ensure such reliability (i.e., to test and assess the reliability of the verification system) it is, in effect, building confidence in its ability to produce reliable confidence assessments. With respect to the personal feelings or assessments of individual inspectors, analysts, or policymakers, however, we are operating at the level of trust—trust in confidence.

Further work should continue to explore both how confidence can be bolstered throughout a verification regime as well as how we measure confidence and further development of indicators. Further work is also suggested to explore what confidence looks like from the "host's" perspective, as the work so far has been mostly focused on inspectors.

Annex A: Measurements as a Tool to Build Confidence in Nuclear Disarmament Verification

Introduction

In the field of nuclear disarmament verification, the term *confidence* has been used in several different contexts (e.g., for the overall nuclear disarmament verification process, for single on-site inspection missions, and for judging measurement results). This annex concentrates on the key elements related to confidence in absence, attribute, and template measurements of nuclear warheads and their associated components.

Absence Measurements

Absence measurements can include larger area surveys, measurements of individual objects such as containers, and combinations of these. Especially large area surveys with multiple objects and limited time require significant prior planning.

When screening larger areas for absence, visual inspection should also be used for selecting priority measurement locations (e.g., containers that could accommodate nuclear warheads or their components). Due to their sensitive nature, data on container shielding materials and masses and on nuclear warhead and component positions within containers will not be available. Although the size of a container provides an upper limit for its shielding potential, considerable uncertainty will remain to what extent radiation signals of weak sources are effectively suppressed. Confidence in detecting weak sources can be increased by:

- Keeping the distance between detector and measurement object as low as possible
- Increasing measurement times
- Using detector types of high geometric and intrinsic detection efficiencies

Setting an Alarm Threshold for Measurements

Some background radiation is associated with every measurement. Its level is location-specific,[9] but even in the same location the number of background counts varies statistically between repeated measurements. In the following, only statistical fluctuations are considered.

The procedure is as follows. First, a background measurement is performed away from any radioactive source. Then, this procedure is repeated with the detector close to the item to be inspected. If the second measurement shows a higher count rate than the background, it has to be estimated whether this result indicates the presence of a radioactive source or is caused by

[9] Background radiation levels and their spatial fluctuations may be particularly high in facilities used for storing and processing nuclear warheads and its components.

statistical background fluctuations. This decision is taken by statistical hypothesis testing with the hypotheses being defined by the two physical options (source vs. background fluctuation).

However, statistical hypothesis tests inadvertently are associated with uncertainties, as a risk exists of accepting a hypothesis although it is wrong. For nuclear disarmament verification, these potential errors are:

1. Rejecting the hypothesis that an elevated count rate is caused by background fluctuations, although it is true, and accepting the alternative hypothesis that it indicates a radiation source, although it is wrong (error type I)

2. Accepting the hypothesis that an elevated count rate reflects background fluctuations, although it is wrong, and rejecting the alternative hypothesis of the presence of a radiation source, although it is true (error type II)

Obviously, option 1 then results in a false alarm, option 2 in ignoring the signals recorded by a non-declared radiation emitter. These two error types are inversely linked (i.e., a reduction in the risk of one always increases the risk of the other). For example, setting a very high alarm threshold will reduce the risk of false alarms—thus providing high confidence that an alarm will indicate a non-declared radiation source, but at the same time will significantly reduce the sensitivity of the measurements—thus reducing the confidence in the results of absence measurements.

How to balance out these two risks is not a scientific issue but depends on the risk perception when interpreting a measurement. In nuclear disarmament verification, apparently both frequent false alarms and low confidence in detecting well-shielded fissile material are challenging.[10]

In the following, these relationships are illustrated by a simple example. Let's assume that a low-tech counter detector is used, that it makes 60 s measurements,[11] and that each measurement has more than 20 counts (allows the use of Gaussian statistics). A first 60 s measurement is performed for recording the background, followed by a 60 s measurement close to the object to be screened. Its potential signal is given by the equation:

$$N_{net} = N_v - N_b \qquad (1)$$

In this equation, N_{net} is the net signal, N_v are the counts from the verification measurement, and N_b are the counts from the background measurement. If additional radiation is not present, the expectation value[12] for N_{net} is zero ($N_v = N_b$). Due to the statistical nature of the radioactive decay

[10] Sometimes this dilemma may be resolved by combining technologies. For example, the absence of excessive shielding inside of a container could be verified by a transmission measurement using a non-sensitive radiation source.

[11] In general practice, longer measurement times will be aimed for in order to increase the sensitivity of the procedure.

[12] It represents the average / mean value in case of a large number of repeated measurements.

process, a normal Gaussian probability distribution is associated to N_{net}. Its standard deviation σ_{net} can be calculated using equation 2:

$$\sigma_{net} = \sqrt{N_v + N_b} \simeq \sqrt{2N_b} \qquad (2)$$

Theoretically there is a 68.3 percent probability that the measurements produce a net signal that is between $-\sigma_{net}$ and σ_{net}. In absence measurements, one is interested in positive N_{net} signals. Their probabilities are given in Table 1. For example, there is a 0.15 percent chance that the background causes N_{net} signals equal or larger than $3\sigma_{net}$ counts. Usually, the alarm threshold for N_{net} is set to a level of X σ_{net} (Table 1), implicitly assuming that hypothesis 1 above (i.e., no source present) is correct. This determines the probability of false alarms as well as the minimum detectable activity if a radioactive source is present.

Table 1: Alarm Threshold Probabilities

Alarm Threshold	False Alarm Probability (%)	Probability (%) of missing a source with activity of 1.5 x background[(a)]	1.0 x background[(a)]
1.00 s_{net} [(b)]	15.85	0.014	0.94
1.64 s_{net}	5.0	0.085	3.44
1.96 s_{net}	2.5	0.18	6.30
2.58 s_{net}	0.5	0.73	14.5
3.00 s_{net}	0.15	1.30	23.9
3.50 s_{net}	0.025	4.01	37.8
4.00 s_{net}	0.0032	8.53	54.0
5.00 s_{net}	0.00003	17.1	82.1

[(a)] For this example, background is assumed to cause a mean number of 30 counts.

[(b)] With N_b=30 a value of $\sigma_{net} \approx 7.75$ results.

A N_{net} signal in excess of the alarm threshold is then interpreted as indicating the presence of a radioactive source. Although the expectation value of the net count becomes greater than zero, if a source is present ($N_v > N_b$), statistical fluctuations can also result in a value below the alarm threshold. Obviously, the risk of such an event depends massively on the source activity, increasing as its difference to the background decreases.

This relationship is illustrated by the example given in Table 1. It has been assumed that the mean number of background counts is 30 and that there are two well-shielded sources with radiation signals of either 150 percent or of 100 percent of the background. Although the risk of not being identified remains low for the stronger source up to a false alarm risk of around one in a thousand, it increases significantly for the weaker source from a false alarm risk of around 5 percent.

The most effective way to improve confidence is through an increase in the cumulative counts. In our example (Table 1), doubling the counts of both background and verification measurement would reduce the risk of misinterpreting the weak source to 2.2 precent for an alarm threshold of 3 σ_{net} and to 34.8 percent for 5 σ_{net}.

To achieve a high level of confidence in radiation measurements, it is essential to estimate the risks of statistically ignoring signals from well-shielded undeclared sources for each application of this technology. Compliance with negotiated error limits will then define the procedures to be followed during inspections. If detector types and minimum source-detector distances are predetermined, measurement times are not subject to discretionary choice. Instead, they are required to allow for the accumulation of the statistically required counts (Table 1).

A similar analysis can be made for a gamma-ray spectrometer that has multiple energy channels. Different radionuclides emit gamma-radiation with their own characteristic energies. By monitoring the number of counts arriving to certain nuclide-specific energy windows of the spectrometer allows one to confirm the absence or presence of that radionuclide of interest. The same formalism as above can be used. The main difference is that now the background, N_b, is determined from the same measurement as N_v. N_v now are the counts in the nuclide specific region of interest (energy window), N_b are the counts in the same width background window next to it.

Impact of Counting Statistics on Confidence

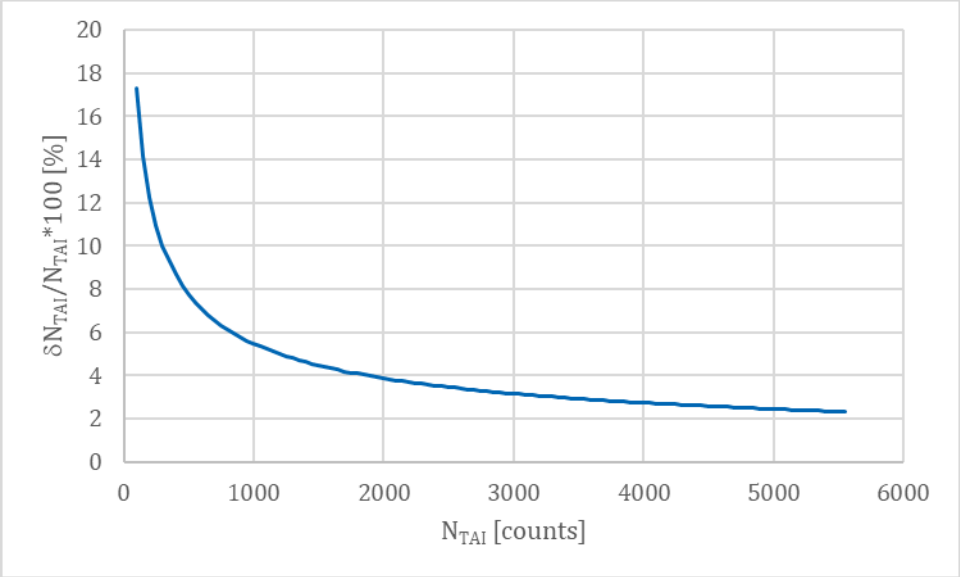
In practice, limiting the error risks discussed above to agreed levels requires extended acquisition times. If the gross counts recorded during a measurement are denoted as M_{TAI} and N_b as its background counts, the net counts emitted by the object (N_{TAI}) are calculated as $N_{TAI} = M_{TAI} – N_b$. Its standard deviation uncertainty is given as:

$$\delta N_{TAI} = \sqrt{M_{TAI} + N_b} \qquad (3)$$

Figure 1 shows how the fractional standard deviation $\delta N_{TAI}/N_{TAI}$ with increasing numbers of N_{TAI} counts, assuming a weak signal identical to background ($N_{TAI} = N_b$). The smaller the relative uncertainties become, the more stringent comparisons between measurements of the same or similar objects can be made, and this improves the associated confidence.

Figure 1 emphasizes the importance of collecting large count numbers, but also that reducing statistical uncertainties by a factor x requires increasing accumulation times by a factor x^2.

Figure 1: The Behavior of the Fractional One Standard Deviation Uncertainty as a Function of TAI Net Counts (NTAI)



Annex B: Bias in Nuclear Disarmament Verification

Biases in Multilateral Verification

Confidence assessments in nuclear disarmament verification will be produced by techniques, equipment, and data using procedures that have been tested and integrated into any future nuclear disarmament agreement. Trust, however, uses a wider set of evidence markers to inform an assessment, some of which may be objective and science-based, some of which may be informed by conscious or unconscious bias.

The Context

The context of multilateral regimes that verify nuclear disarmament agreements is characterized by several factors:

- Assessments are derived from repeated events taking place over the course of years, if not decades; results must be consistent with previous results and/or expectations in accordance with treaty obligations.
- Assessments can include qualitative elements (such as whether the host is cooperative).
- Humans conduct assessments and humans are fallible.
- Humans have unconscious biases that they may not be aware of that may or may not be relevant to the confidence assessment.
- Compliance judgments rely heavily on data, but can be affected by the "gut feeling" of individual decision makers.

These factors will influence both trust and confidence.

The Human Factor

Although new and emerging technologies hold promise in international security and specifically may add value to elements of the verification process, ultimate assessments of compliance with a disarmament agreement require a human in the loop. Explorations of the role of technology in eliminating, or perpetuating bias, are beyond the scope of this paper.[13] Humans can perceive behavior that is non-quantifiable, consider context, and weigh the importance of factors that may influence hard data. Perhaps the greatest example of the necessity of human reasoning in international security can be drawn from the Soviet "false alarm" incident of 1983.[14]

[13] For an introduction to algorithmic bias, see United Nations Institute for Disarmament Research, *Algorithmic Bias and the Weaponization of Increasingly Autonomous Technologies* (UNDIR, 2018), https://unidir.org/wp-content/uploads/2023/05/algorithmic-bias-and-the-weaponization-of-increasingly-autonomous-technologies-en-720.pdf.

[14] The "false alarm" incident refers to the malfunction of a Soviet missile detection system that incorrectly alerted Soviet officers that the United States had launched five ICBMs toward the Soviet Union. The lieutenant colonel on

In a verification scenario, humans are not only *able* to take factors into consideration that a machine may not perceive, but are often *incapable* of leaving said perceptions out of their internal deliberations.[15] These perceptions can lead to bias, whether conscious or unconscious. Bias can be defined as prejudice in favor of or against one thing, or group compared with another, usually in a way considered to be unfair. *Conscious bias* is the awareness of those feelings, emotions, position, and underlying values. *Unconscious bias* is the inactive or unaddressed beliefs, values, and opinions that affect how we interact with the world and one another. Unconscious bias can lead to a misperception of impartiality, where individuals believe they are being impartial when in fact they are making decisions based on factors that are not relevant to the situation. Many potential types of human factor bias can be identified that may influence the nuclear disarmament verification process, including on the part of the inspectors and the host inspected as well as the dynamics within an inspection team.

Discussion of all forms of bias that may be present in a verification scenario is beyond the scope of this paper. For the IPNDV's purposes, it is merely important to recognize that bias may be present:

- Within the inspecting team, leading to questioning authority and/or expertise in assessments
- Against the host party, leading to concerns about compliance with nuclear disarmament agreements
- Against the inspecting party, leading to complaints about impartiality and the manner in which compliance is being assessed.

Although all humans experience biased decision making, beliefs about race, gender, ethnicity, nationality, or socioeconomic status can perpetuate bias in a dangerous manner. These beliefs are more likely to manifest in homogenous groups of people in a workplace, government, or culture. Among all the benefits of diverse teams,[16] unconscious bias about certain characteristics

---

duty, Stanislov Petrov, had only moments to decide whether he would report the notification to his superiors, undoubtedly triggering a series of events that would lead to a counterattack. Petrov was uniquely familiar with the Soviet's missile detection system, Oko, and was well aware that false alarms could occur. Furthermore, Petrov later recounted that he believed an unprompted American attack would consist of dozens of ICBMs, if not more, rather than simply five. Petrov ultimately reported that the alarm was false and he would be proved correct. Months later, the Soviets discovered that Oko had incorrectly interpreted reflections under odd atmospheric conditions in the United States to be ICBM launches. Brian J. Morra, " The Near Nuclear War of 1983," *Air & Space Forces Magazine*, December 2, 2022, https://www.airandspaceforces.com/article/the-near-nuclear-war-of-1983/.

[15] Humans can process up to 11 million bits of information per second, but are only capable of consciously taking in about 40–50 bits of information per second. The barrage of information prompts humans to take shortcuts and make decisions without necessarily understanding the process of how said decisions are made. Emily Kwong, "Understanding Unconscious Bias," NPR, July 15, 2020, https://www.npr.org/2020/07/14/891140598/understanding-unconscious-bias.

[16] Diverse teams are proven to be more creative, innovative, and effective at problem solving. This is likely due to the diminished role of groupthink, which describes the phenomenon of homogenous groups reinforcing a singular view due to their desire for conformity within the group. Groupthink is dangerous when the collective disregards

like skin color or gender can be confronted as judgments about said characteristics and can be proven false through continued interactions with diverse individuals. One of the greatest tools to combat bias is to prioritize diversity within a group and collectively overcome the biases that emerge through myriad strategies, which will be outlined in the conclusion.

Why Bias Matters for Nuclear Disarmament Verification

Bias has the potential to impact the effectiveness, efficiency, and credibility of the implementation of a multilateral nuclear disarmament verification regime. Verification is a necessary component of a nuclear-weapons-free world, and questions relating to the credibility of assessments of compliance have the potential to delay or derail nuclear disarmament agreements. More evidence is required to assess the role bias has played in past inspections and how it could be present in future inspections.

Anecdotally, several IPNDV participants also recognized the challenge of *positive bias* influencing inspections. Primarily in the context of long-term inspection regimes (10+ years), the continuity of inspectors over the entire regime was identified as possibly contributing to a more collegial yet unofficial environment. This environment has the potential to be intentionally or unintentionally exploited, as collegial inspectors can be more inclined to excuse minor variances or infractions, according to those with first-hand experience as inspectors. The impact of positive bias is also underexplored as a topic and requires further evidence-gathering and analysis.

The IPNDV has a responsibility to extensively consider all factors that may complicate the achievement of global zero from a verification viewpoint. Decades of technical and scientific research have influenced and improved nearly every aspect of the IPNDV's proposed verification regime. It is only logical that the partnership also invests time and resources into better understanding the challenges associated with the authoritative role of human trust and confidence in a verification regime.

Conclusion

Moving forward, the IPNDV may wish to explore the following strategies for recognizing and addressing conscious and unconscious bias in nuclear disarmament verification:

- Commission a study on the effects of bias in past verification regimes
- Conduct further assessments of the role of confirmation bias and the prominence effect
- Invite guest speakers with experience in multilateral verification regimes to speak to their experience with bias
- Engage a trained professional to conduct training sessions for the IPNDV on overcoming bias in the workplace, in order to better equip future inspectors

other views, especially those that have more logical backing than the chosen view. "Groupthink," *Psychology Today*, https://www.psychologytoday.com/ca/basics/groupthink

- Develop a toolkit for addressing bias before, during, and after an inspection and integrate this toolkit into future tabletop exercises, or verification games alongside other PPTTs being tested
- Capture the outcomes of the above in a final IPNDV deliverable to inform future diplomats of the measures they can draw from when setting the parameters for developing a disarmament inspectorate

About IPNDV the International Partnership for Nuclear Disarmament Verification

The International Partnership for Nuclear Disarmament Verification (IPNDV) convenes countries with and without nuclear weapons to identify challenges associated with nuclear disarmament verification and develop potential procedures and technologies to address those challenges. The IPNDV was founded in 2014 by the U.S. Department of State and the Nuclear Threat Initiative. Learn more at www.ipndv.org.