SAND2016-9371C

Developing a System Evaluation Methodology for a Warhead Monitoring System

Cliff Chen[1], Crystal Dale[2], Sharon DeLand[3], Angela Waterworth[4], Tom Edmunds[1], Doug Keating[2], Matthew Oster[4]

[1]Lawrence Livermore National Laboratory, Livermore, CA 94550
[2]Los Alamos National Laboratory, Los Alamos, NM 87545
[3]Sandia National Laboratories, Albuquerque, NM 87123
[4]Pacific Northwest National Laboratory, Richland, WA 99352

1. Introduction

Past agreements limiting the numbers of nuclear weapons focused on nuclear weapon delivery systems and the warheads deployed on these systems. As the number of nuclear weapons decreases, future agreements may require the monitoring of nuclear warheads throughout greater portions of the weapons enterprise. Such agreements would pose a new set of verification challenges that may require new monitoring approaches, technologies, and procedures. The monitoring system for such agreements may be more intrusive than in past agreements, particularly as additional elements of the warhead lifecycle are potentially impacted, including deployment, storage, maintenance, transportation, and dismantlement, thereby creating significant operational, safety, and security challenges for a host country.

Since the terms of future agreements are subject to negotiation and hence unpredictable, it is important to develop a range of technical solutions for potential warhead monitoring systems, as well as the capability to assess their effectiveness and explore tradeoffs. This paper describes the initial development of a system evaluation methodology for quantifying the system performance of a monitoring approach. The evaluation methodology includes (1) an evaluation framework that includes defining monitoring objectives, functional architectures and associated evaluation criteria, and evaluation scenarios, and (2) evaluation tools such as analysis algorithms and simulation tools for quantifying system performance. The basic tenets of this approach have been used by multiple communities [1][2] but have not been as well developed for arms control verification. When fully developed such a capability would support the technical community in designing and assessing monitoring approaches, inform the policy community about the performance and impacts of monitoring options, and potentially help guide future R&D investments for warhead verification and monitoring.

2. Notional Treaty Provisions and Methodology Assumptions

The evaluation methodology assumes a proposed treaty that governs the limitation and reduction of all nuclear warheads. We assume a New START-like monitoring regime is in place for deployed warheads; the warhead monitoring system being assessed governs non-deployed warheads. All warheads are assumed to be of a single type. Deployed, non-strategic warheads are not considered. We also assume

Work performed under the auspices of U.S. Department of Energy by LLNL under Contract DE-AC52- 07NA27344. LLNL-CONF-695000

that each TAI is containerized. Each party periodically exchanges data that includes the quantity and location of each Treaty Accountable Item (TAI). Each storage, deployment, or maintenance/production facility will be subject to some type of inspection.

This evaluation methodology is being developed for a cooperative monitoring system that monitors declared facilities. The detection of undeclared facilities is an important part of an overall verification regime, but better addressed by other means.

3. Evaluation Methodology

Figure 1 depicts the components of the evaluation methodology. The evaluation framework provides the structure within which a monitoring approach is evaluated. It consists of: (1) establishment of monitoring objectives, (2) a functional decomposition of the monitoring system, (3) quantitative evaluation criteria, and (4) evaluation scenarios. The monitoring approach to be evaluated must the set of monitoring functions, process diagrams describing how they fit together, (5) decision trees that describe an inspector's logic process, and functional performance data. To perform the evaluation, evaluation tools include sets of (6) analysis algorithms and (7) a Discrete Event Simulation capability. The methodology can be used to quantitatively assess system performance, help identify the value of specific monitoring functions, and compare the performance of implementation options. The components highlighted in red and enumerated above are described in this paper.

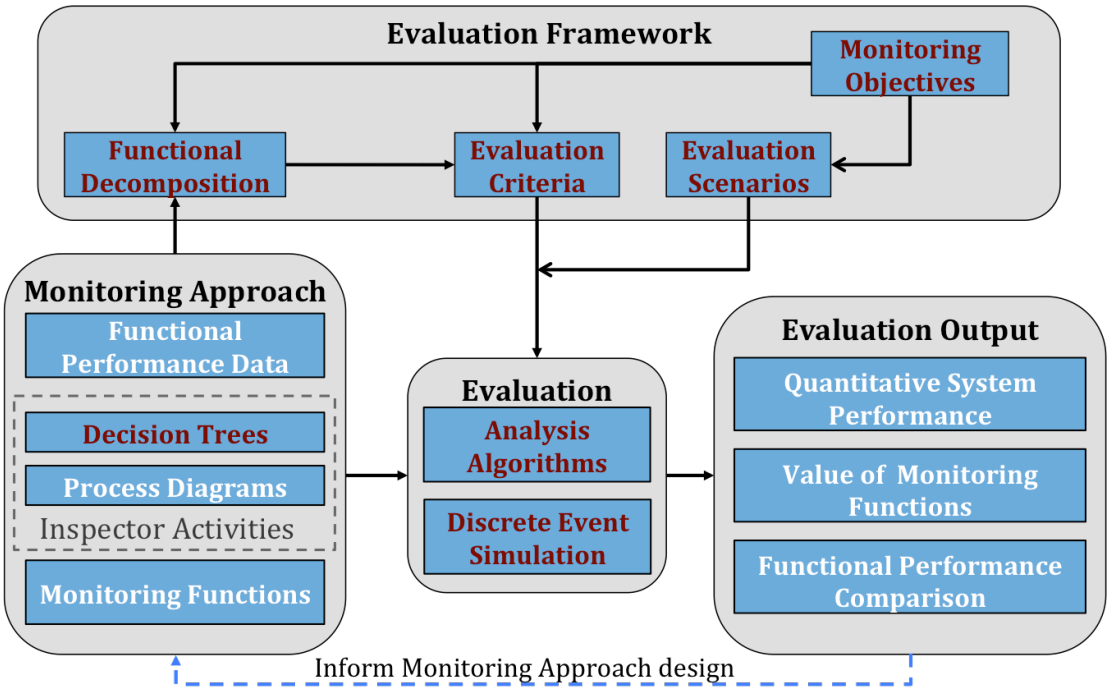

Figure 1. Evaluation Methodology Structure

Monitoring Objectives

The performance of a monitoring system is evaluated against its ability to provide confidence that a treaty partner's declarations are accurate and complete.

Confidence in the declaration can have multiple meanings, therefore treaty verification objectives must be agreed upon and specified in order to design and evaluate a monitoring system. For our purposes, we used three verification objectives. These are to confirm the declaration, including:

1. The number of Treaty Accountable Items (TAIs),
2. That the TAIs are warheads, and
3. The dismantlement of warheads

The first two objectives highlight potentially distinct policy objectives. A treaty partner may want to know (as a worst-case scenario) that the declaration is complete and that there are not more warheads than declared. A secondary objective may be to know that the TAIs under monitoring are warheads, so that opportunities for strategic mis-calculus are minimized. In specifying these objectives it is important to distinguish between high-level verification objectives and supporting activities.

Once verification objectives are defined, monitoring objectives can be specified for the cooperative monitoring system. These are distinct from, but closely associated with the verification objectives. The monitoring objectives[1] specify what the cooperative monitoring system is supposed to do. For the verification objectives above, we defined the corresponding monitoring objectives:

1. Confirm the declared number of TAIs at all locations within the declared enterprise
2. Confirm that TAIs in the declared enterprise are warheads
3. Confirm the TAI to be dismantled is a warhead, and the dismantlement of that warhead

There are other potential monitoring objectives that could be specified. The monitoring objectives for the cooperative system may also depend on the information needed to support other monitoring capabilities such as National Technical Means and open-source analysis. For example, confirming that the TAIs in the declared enterprise are warheads may support confidence in the completeness of the declared number of TAIs, particularly as related to confidence in baseline numbers and the lack of hidden stockpiles. Another objective may be to demonstrate that declared facilities are not being used for the maintenance or production of undeclared TAIs. For the purposes of this paper, we will focus on the three monitoring objectives defined above.

Functional Decomposition

A functional decomposition is an architecture that describes the relationship between the monitoring objectives and the monitoring system functions and sub-functions. Some sub-functions represent technology implementation options for a

[1] In this paper monitoring is used for cooperative monitoring. In other contexts, monitoring refers to the totality of data collection activities, including NTM.

given function. This architecture helps structure the diverse array of monitoring functions and assists in the design and comparison of monitoring approaches. A notional functional decomposition is shown in Figure 2.

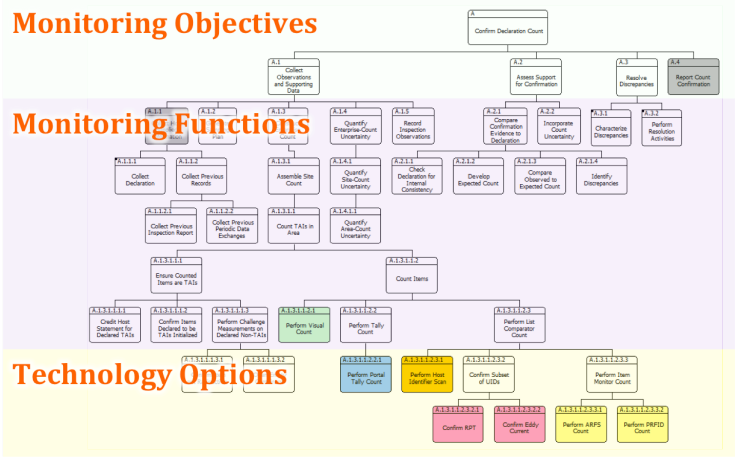

Figure 2. A functional decomposition defines how monitoring functions support objectives and how functions are implemented by technology options.

For example, the function of counting TAIs is comprised of counting TAIs at a site through unattended monitoring and inspections, and correlating information at the enterprise level. Site-level counting includes both counting items in designated areas and ensuring the counted items are TAIs (which may mean crediting a host's statement that an item is a TAI subject to monitoring). Counting items could be performed through a combination of visual counting by inspectors, list comparator counting that confirm a host or inspector's human-readable unique identifier or electronic item monitor, or tally counting such as with a radiation portal monitor.

The hierarchical structure of the functional decomposition creates traceability between the functions performed by the monitoring system and the monitoring objectives they support. It also supports the definition of evaluation criteria for each function and facilitates the direct comparison of implementation options.

Evaluation Criteria

Defining evaluation criteria that accurately reflect objectives is a fundamental challenge in any evaluation activity. While the monitoring objectives are to confirm the declarations, defining confidence in that confirmation is challenging as one could imagine any number of scenarios not well reflected by a given criterion. In many cases, it is more intuitive to define system performance through the inverse problem, by defining the monitoring system as a detection system. In this construct, the goal of a monitoring system is the timely detection of discrepancies (associated with the monitoring objectives).

The International Atomic Energy Agency, for example, has specific definitions for timeliness goals and significant quantities for safeguards planning that are derived from technical analysis. In arms control these concepts are more state-specific, and correlate with the timely detection of activities contravening treaty limitations at levels that introduce instability into the strategic relationship. Without getting the details of defining these concepts or quantitative requirements, we can still characterize monitoring system performance against a range of conditions. For

example, one possible set of metrics for monitoring the declared count is:

- Time to first detection of a discrepancy as a function of X ≥ 0 discrepancies
- Time to confidence in the magnitude of the discrepancy

"Confirm that TAIs in the declared enterprise are warheads" consists of statistical sampling, Chain-of-Custody, and warhead confirmation measurement activities. The high-level metrics are similar to those above – time to first detection and time to confidence in the discrepancy magnitude as a function of for X ≥ 0 non-warheads.

At a lower level of the hierarchy, an important set of criteria is the performance characterization of individual warhead confirmation technologies. If warhead confirmation technologies are part of a future treaty regime, the treaty will need to specify a warhead definition, whether as a set of attributes or against a set of reference templates. However, this negotiated definition will be informed by assessments of both what is needed from a policy perspective, and what is possible from a technical perspective.

What characteristics constitute a sufficient warhead definition has historically been discussed at length, and a number of efforts are currently assessing novel approaches to a warhead definition. In order to focus on the technical performance evaluation, we defined the performance criterion as the ability of a measurement to distinguish between a reference object and a set of test objects. These test objects may represent other objects in the enterprise, other possible warhead definitions, or potential spoofs. As each technology is sensitive to different warhead characteristics, they will likely have varying performance against different test objects. A notional representation of this criterion is represented in Figure 3.

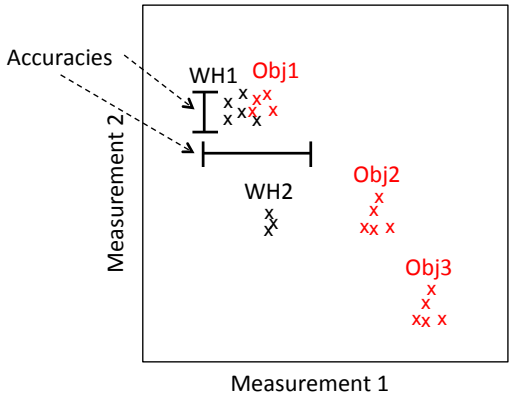

Figure 3. Notional representiation of two measurements used to distinguish between a reference object and test objects.

The specific accuracy metric can be defined as set of receiver operating characteristic (ROC) curves for the comparison of each test object to the reference. In this classification approach, an adjustable discrimination threshold impacts the true positive (TP), and false positive (FP) rates, which reflect the accuracy of correctly accepting the reference object and incorrectly accepting the test object, respectively. Their dependence can be mapped out on the ROC curve. For simplicity, points on the ROC curve can be shown as a table of FP rates for a specified TP rate. Table 1 depicts notional technology performance as false positive rates for a set of reference objects for a true positive rate of 99%.

TP = 99%	Object A	Object B	Object C
FP [60 min]	60%	92%	77%
FP [120 min]	15%	79%	4%

Table 1. Notional technology performance for distinguishing reference and test objects.

There are two components to assessing the technology performance: performance under stochastic uncertainty, and the sensitivity to nuisance parameters. Stochastic performance is illustrated in Table 1, showing how longer measurement times result in lower FP rates and therefore better discrimination. Nuisance parameters represent the operational, environmental, and detector specific conditions that result in systematic uncertainty. For example, some technologies may be much more sensitive to alignment or the position of a TAI within its container. It is important to assess these sensitivities, evaluate their impact on classification performance, and potentially identify procedures that mitigate the performance degradation.

Evaluation Scenarios

Evaluation scenarios outline the conditions under which the monitoring system is expected to perform. The evaluation scenarios are derived from verification concerns, and the evaluation criteria are evaluated within the context of the scenarios. They serve to specify the types and patterns of discrepancies that may be encountered. For example, some simple scenarios related to the declared count include:

- TAI inventories are consistent with the declaration
- Excess TAIs present in declared facilities are undeclared and uninitialized
- Excess TAIs are declared upon site inspection, but declared to be from legitimate transport or deployment exchange processes
- TAIs declared to be on-site are missing
- TAIs declared in periodic data exchanges never happen to be present during site inspections

It is important to develop a comprehensive set of scenarios that reflect verification concerns and potential vulnerabilities. The evaluation criteria can then be assessed against its ability to perform under each of these scenarios.

In addition, Table 2 describes three levels of realism under which to assess monitoring system performance. At higher levels of realism, the evaluation criteria are slightly modified such that the monitoring system must provide detection confidence above an expected level of functional false positives or honest enterprise errors (e.g., paperwork errors, handling errors).

Different types of enterprise errors may have different impacts on system performance and may help define requirements for additional monitoring functions. For example, in New Start, an inspector has the ability to use Radiation Detection Equipment (RDE) to confirm that a shrouded object is not radioactive and therefore should not be counted as a warhead. This ability to perform a negative warhead

confirmation measurement (confirming an item is not a warhead) may help resolve some potential discrepancies during an inspection.

Enterprise Errors	Functional performance	Description
None	Ideal	*Artificial construct for framing and testing metrics*
None	Realistic	*Monitoring system provides detection confidence above expected functional false positives*
Realistic	Realistic	*Monitoring system provides detection confidence above background of honest enterprise errors*

Table 2. Levels of realism for assessing monitoring system performance.

Decision Trees

Decision trees define the inspector logic process in response to monitoring activity observations. While these decision trees are part of the monitoring approach specification rather than the evaluation process, they are an iterative part of the evaluation methodology, integrating the evaluation scenarios and analysis algorithms. A notional decision tree is shown in Figure 4. It walks through the steps in the inspection process, including inspection planning, the performance of monitoring functions, potential follow-on activities, and the identification of detection events. The decision trees support the identification of monitoring gaps and are part of the analysis logic for quantifying the evaluation criteria.

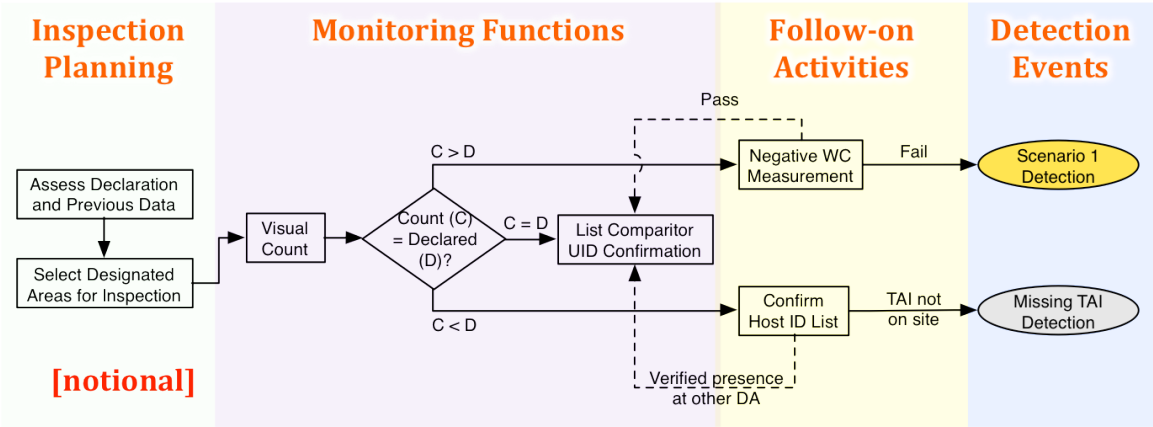

Figure 4. Decision trees define the inspector logic process in response to monitoring activity observations. They support the identification of monitoring gaps and are part of the analysis logic for quantifying the evaluation criteria.

A decision tree needs to be developed for each proposed monitoring approach. While aspects of it may be modularized and portable between approaches, the consistency of the logic must be carefully considered.

Evaluation criteria and decision trees for dismantlement confirmation are described

in a concurrent paper by Dale, et. al. [3].

Analysis Algorithms

Monitoring data is assessed in such a way as to try to form a consistent picture across all sources of information about whether a state's activities are consistent with its declarations. One of the challenges in developing analysis algorithms for a future treaty regime is understanding the potential range of state activities and how to assess whether the declarations, technical monitoring data, and inspection results are consistent with them. This "inverse problem" is difficult to solve without the constraints of a given framework.

To make progress in algorithm development and metrics quantification, we have framed the monitoring system as a detection system. For each specified evaluation scenario, the evaluation criteria described above are the time to first detection and time to confidence in the discrepancy magnitude. Under this framework, algorithms may be developed to quantify these metrics.

The requirements on these algorithms also depend on the evaluation scenario being considered. For example, in a simple detection scenario, standard approaches such as likelihood estimates and parameter estimation may be sufficient. More complex scenarios, such as those that involve correlated activities between inspection sites and differences in declaration timing, require more complex algorithms that consider the consistency of declarations across time and between sites.

Statistical methods may be sufficient for quantifying the detection of discrepancies arising from honest host mistakes, which may have some level of random distribution. However, an effective monitoring system must provide confidence in the context where a party may engage in a range of cheating behavior. The other party would be aware of the logistics and limitations of the verification regime and engineer a strategy that minimized the probability of detection. Under these conditions, game theoretic approaches are necessary for quantifying performance. A concurrent paper by Edmunds and Chen surveys some potential approaches [4].

A more rigorous assessment extended beyond the detection system framework would consider the range of potential scenarios and incorporate algorithms that interpret the monitoring data and provide uncertainties with regards to scenario consistency. Baysian-based "situational awareness" frameworks have been used to assess nuclear terrorism scenarios [5] and may be helpful in this context.

Discrete Event Simulation

Analytic statistical methods are complicated by the fact that monitoring functions are performed across a dynamic nuclear enterprise that encompasses random and scheduled processes. A discrete event simulation (DES) capability enables the quantitative assessment of system performance by modeling the underlying dynamics of the TAI lifecycle and the monitoring system. The DES simulations enterprise processes and frequencies, and the statistics associated with inspections and the stochastic performance of monitoring activities.

The DES consists of three primary modules: the TAI module, which models

enterprise processes, the declaration module, which generates periodic data exchanges and on-site declarations consistent with a particular monitoring approach, and the monitoring module, which models all of the monitoring functions, including unattended monitoring activities and inspections. A simplified picture of the architecture is shown in Figure 5.

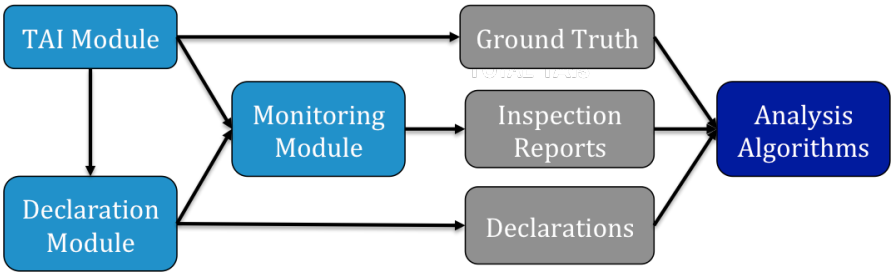

Figure 5. Simplified architecture of the DES. The DES models the interface between the dynamic nuclear enterprise and the monitoring system.

The DES is initialized with the specific evaluation scenario that is being assessed. It outputs a ground truth, describing what actually happened in the enterprise, a series of declarations, and a set of inspection reports that describe the outcome of the inspections (we currently assume unattended monitoring data is only accessible during inspections and not transmitted in real time). Analysis algorithms use these outputs to quantify the evaluation criteria. The DES capability is further described in a paper by Oster, et. al. [6].

4. Example Analysis

To demonstrate the use of the methodology, we need to define a notional monitoring approach and nuclear enterprise for evaluation.

Notional Monitoring Approach

We extrapolate a monitoring approach from current regimes involving identifiers and random inspections. The monitoring approach allows for 12 random inspections per year. The host is required to assign an alphanumeric identifier to each TAI container (hereafter referred to as just the TAI). During an initialization phase of the treaty, each TAI is initialized by applying an additional unique identifier (UID) to the TAI, along with container seals and tamper indicators. The unique identifiers are recorded in a dual-controlled, secure database. It is expected that routine enterprise activities such as maintenance or deployment exchange will result in breaking the container seals. These TAIs will be declared and reinitialized in the presence of an inspector. Periodic data exchanges are provided every 6 months, with a 30-day security delay, and include the identity and location of each TAI in the enterprise, along with deployment exchange, maintenance, and shipping records.

At each inspection, the host provides the inspection team an up-to-date on-site declaration for that site only. The inspectors select a subset of monitored areas for inspection, where they can perform a visual count, confirm the UID of a sample of the TAIs, and check seal/container integrity. The inspectors choose the monitored

areas with a set of inspection rules that include: (1) monitored areas with uninitialized TAIs are inspected, (2) at least 30% of monitored areas are inspected, and (3) at least one monitored area must be chosen at random. After inspecting the monitored areas, the inspectors then perform warhead confirmation measurements on a small sample of TAIs that are brought to a designated measurement station.

Nuclear Enterprise

We assume that there are 2000 TAIs, half deployed and half in storage. The notional enterprise is comprised of 8 military bases consisting of deployment, storage, and on-site maintenance facilities. Processes at the bases include deployment exchange, on-site maintenance for the exchange of limited life components, shipping/receiving to the designated maintenance site for extended maintenance or dismantlement, and deployment drills. Each base is assumed to have 10 storage areas, with 100 deployed and 80 non-deployed TAIs per base (on average).

The enterprise has a designated maintenance/dismantlement site where TAIs undergo extended maintenance. The site is comprised of 10 storage areas, a set of maintenance facilities, and a single dismantlement facility. On average, 200 TAIs are at the site. Processes include shipping/receiving to military bases, temporary storage awaiting maintenance/dismantlement, and maintenance/dismantlement activities.

Example Analysis

In this simple example, we assess a scenario where undeclared TAIs are randomly introduced across the enterprise immediately after the initialization phase of the treaty. The primary monitoring function that detects these discrepancies is the visual count of the monitored area. The statistics are primarily driven by a stratified random sampling approach of selecting sites and monitored areas. The monitoring area selection rules, the frequency of TAI processes and movements, and the initialization of any undeclared TAIs that are found complicate the statistics.

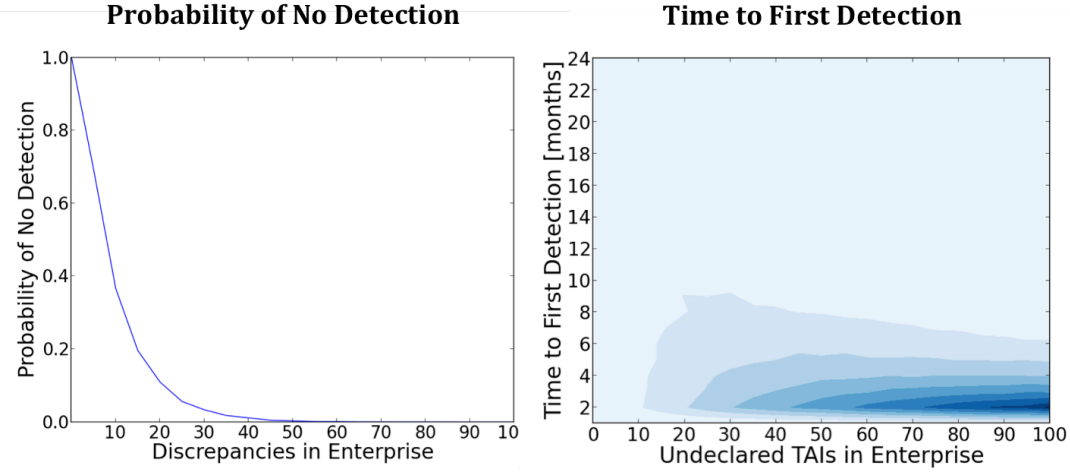

Figure 6. Probability of no detections of undeclared TAIs in 24 months and the distribution of time to first detection as a function of the number of discrepancies introduced.

The scenario was modeled in the DES for a period of 24 months, introducing up to 100 excess TAIs through the enterprise.

The left plot in Figure 6 shows the probability of no detected discrepancies in 24 months. The right plot depicts the probability density of the time to first detection of a discrepancy. Even low numbers of excess undeclared TAIs are likely to be detected within the first few inspections. Time is analogous to the number of inspections at a high level, but also incorporates the probability of an inspection.

A simple analysis algorithm is a parameter estimation of the number of discrepancies in the enterprise as a function of the cumulative discrepancies detected. The DES is used to calculate a 3-D probability distribution as a function of discrepancies introduced, time, and cumulative discrepancies detected. This is shown in Figure 7. At early times the probability distribution is vertical, since due to low sampling statistics, the low numbers of detected discrepancies result in large uncertainties in the estimate number of undeclared TAIs. At later times, with more samples, the relationship has a strong linear correlation.

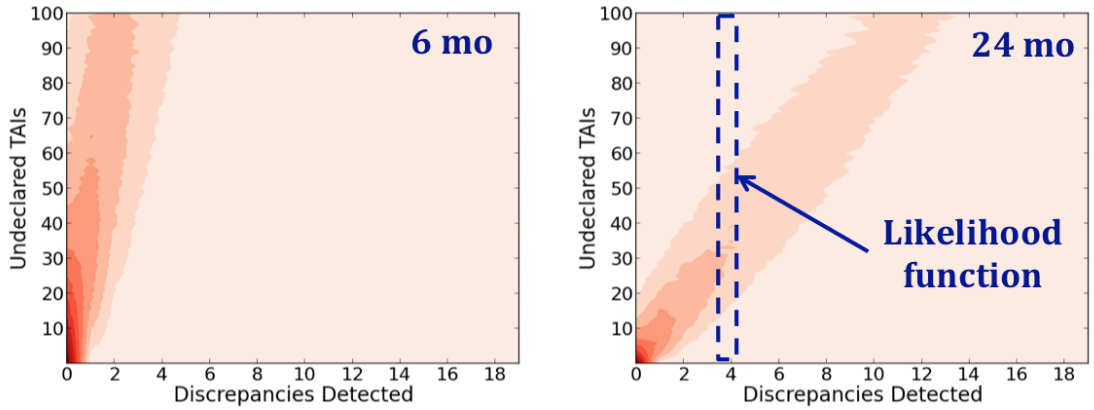

Figure 7. The probability density of the cumulative number of detected discrepancies at 6 and 24 months as a function of the number of undeclared TAIs introduced. A family of likelihood functions can be calculated for each time and number of discrepancies detected.

A family of likelihood functions can be calculated for each time and cumulative number of detected discrepancies. These likelihood functions can then be used to estimate the number of undeclared TAIs across the enterprise for a specified detection pattern (a given instantiation of the simulation) as seen in Figure 8. In this simulation, 56 undeclared TAIs were present. Early inspections did not find any discrepancies, resulting in an estimation of zero discrepancies in the enterprise. The error bars are calculated from the likelihood functions. The estimates start to converge on the actual number after about 12 months.

Note that this analysis is shown as an example only. The robustness of the probability distribution for cumulative detections has not been tested for its sensitivity to enterprise dynamics.

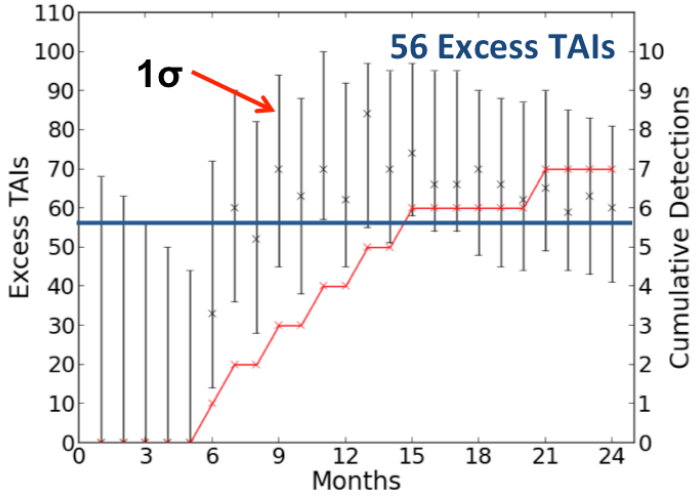

Figure 8. Parameter estimation of the number of excess TAIs across the enterprise for a specific detection pattern.

5. Summary and Future Development

This paper described the initial development of a system evaluation methodology for quantifying the system performance of a treaty monitoring system. The evaluation framework and evaluation tools described aid in the quantitative assessment of monitoring system performance.

Significant work remains in to further develop the evaluation methodology. A deeper assessment of dishonest scenarios may require game-theoretic approaches for metrics quantification. In addition, the functional architecture may need to be expanded to include assessments of the security and integrity of the monitoring data along its entire chain, including generation, transmission, data management, and inspector analysis.

Analysis frameworks for analyzing monitoring data consistency with a range of scenarios, such as Baysian-based "situational awareness" frameworks, will provide a more comprehensive assessment of the monitoring system capabilities. Future algorithms may also need to assess the uncertainty in understanding the enterprise dynamics of the monitored party. The degree to which this is important, and whether algorithms must incorporate convergence on detection statistics and enterprise dynamics will need to be explored.

The evaluation methodology has also not directly addressed the important topics of authentication/certification of technologies, confidence in the template, human factors, and confidence gained through provenance. Whether these elements can be incorporated into the evaluation methodology described needs to be considered.

For any treaty verification regime, the performance of a monitoring system is only one consideration in negotiating a verification regime. Other factors, such as the information put at risk by the regime and its impact on enterprise operations, are often as, if not more important, than system performance. These factors also need to be systematically considered and evaluated against performance tradeoffs. The evaluation methodology described may be able to assist those assessments.

6. References

[1] "Systems Engineering Fundamentals", Section 5, January 2001, prepared by The Defense Acquisition University Press, Fort Belvoir, Virginia 22060-5565.

[2] "What is Operations Research?", Institute for Operations Research and Management Science, June 5, 2015.

[3] "Evaluation Methodology for Dismantlement Verification", C.B. Dale, D. Keating. B. Okhuysen. *57th Annual Meeting of the Institute for Nuclear Material Management*, Atlanta, GA (2016).

[4] "Statistical Sampling Methods for Treaty Verification in Dynamic Environments", T.A. Edmunds and C.D. Chen, *57th Annual Meeting of the Institute for Nuclear Material Management*, Atlanta, GA (2016).

[5] "Exploitation of Ambiguous Cues to Infer Terrorist Activity", K. Ni, D. Faissol, T. Edmunds, R. Wheeler, *Decision Analysis*, Vol. 10, No. 1, pp. 42-62, March 2013.

[6] "Using Simulation to Support Monitoring System Design and Evaluation", M. Oster, A. Waterworth, D. Keating, C.B. Dale, S.M. DeLand. *57th Annual Meeting of the Institute for Nuclear Material Management*, Atlanta, GA (2016).