CRITERIA FOR COMPARATIVE EVALUATIONS OF TREATY VERIFICATION MONITORING SYSTEMS

Angela M. Waterworth and Jacob Benz, Pacific Northwest National Laboratory, 902 Battelle Boulevard, P.O. Box 999 (K7-20) Richland WA 99352, (509)375-3839
angela.waterworth@pnnl.gov

ABSTRACT

When considering an arms control verification regime, there may be many combinations of technologies, data, and procedures to evaluate. Analyzing alternatives is made more complex by competing and conflicting interests of treaty parties. In 2017, a team from the U.S. Department of Energy national laboratory complex developed criteria to assess monitoring systems implemented for treaty verification. The criteria are independent, collectively exhaustive, measurable, and reflective of the distinct host and monitor perspectives. The team also developed a method to score alternatives that could be used to inform decisions on system design. In this paper, we review the criteria and evaluation process.

Keywords: Decision Making, Criteria, Analytical Hierarchy Process, Treaty Verification, National Security

INTRODUCTION

Analyzing options for a monitoring system, including options for technologies, data, and procedures to meet a specific treaty verification scenario, can be complicated and difficult. Competing interests lead to a multitude of tradeoffs, especially considering the competing and contradictory perspectives of host and monitor – roles that each country may assume at different times throughout the regime. While acting as the host, the objective is to demonstrate compliance while preventing disclosure of information not covered by the treaty; the monitoring party's objective is to conduct the inspection in order to determine compliance. Evaluation criteria may offer a tool to understand and evaluate competing options. Effective criteria will identify and prioritize factors of importance considering the host and monitor perspectives and should be unambiguous, independent from each other, collectively exhaustive, and measurable (whether this is quantitative or by a systematic subjective evaluation). When used in the evaluation of options, criteria help to create a consistent framework that can assist in selecting the optimal solution among a set of competing alternatives.

For this study, we endeavored to enhance a previous set of criteria to systematically evaluate chain of custody-based monitoring systems used to support treaty verification objectives within a hypothetical arms control verification regime in which the dismantlement of nuclear weapons is monitored within a "plant within a plant" facility at a nuclear weapons assembly and disassembly facility. This study also considered other evaluation criteria, including those developed for the Portal Monitor for Authentication and Certification (PMAC) project, to provide a unified and generalized set of criteria. In our analysis, we consider a monitoring system to include the data, technologies, and procedures that combine to create a monitoring system.

EVALUATION CRITERIA

The evaluation criteria are organized according to the following hierarchy:

- **Values** represent what is important or of greatest interest from the host and monitor perspectives, classified as either benefits that support the Host or Monitor in achieving their interests or costs associated with implementing the regime
- **Objectives** capture what the Host and Monitor want to achieve
- **Measures** are observable features or characteristics; the efficacy of a monitoring system in supporting the objectives and values can be evaluated by examining the features and characteristics of the monitoring system
- **Metrics** are measureable factors related to the observables that enable the objective assessment of the criteria. Ultimately, all metrics will be quantifiable; while some can be quantified presently. Further research is needed to develop approaches to evaluate other metrics, such as those related to the ability to detect or the likelihood of detecting tampering. Metrics are traceable through the criteria and objectives to Host or Monitor values.

This effort presents two distinct sets of evaluation criteria: one to evaluate the monitoring regime from the perspective of the monitor and a second from the perspective of the host. While the project team planned to develop a single set of criteria, the unified criteria sufficiently represent the distinct considerations from the perspectives of the host and monitor in evaluating a monitoring regime. Using two sets of criteria creates two separate assessments for each monitoring system option, which may make it more difficult to make a decision between options. However, the criteria as presented reflect the tension between these perspectives more realistically.

Additionally, these criteria are structured according to benefit and cost: criteria are aligned into a family of criteria for benefit. This approach produced criteria that are appropriate, useful, mutually exclusive, and represent the distinct and, at times, conflicting interests of both the Host and Monitor. The benefit and cost framework, which serves as the foundation for these criteria, supports the evaluation of both the effectiveness and efficiency of implementation options from both the Host and Monitor perspectives. Finally, elimination of solutions from the criteria reduces bias and prejudice and ensures a fair evaluation of all implementation options.
The criteria for the Host and Monitor are presented in Figures 1 and 2.

EVALUATION PROCESS

This effort considered multi-criteria decision analysis approaches in two major categories: an absolute approach and a comparative approach. The absolute approach produces a benchmarked score for a monitoring system that stands alone. If a monitoring system receives a (nominal) score of 50, it is possible to compare that score to the scores of historical systems and consider marginal improvements in the score due to technological or procedural improvements. The comparative approach produces a score for each monitoring system that indicates its performance relative to the other systems that were included in the comparison.

The effective use of an absolute evaluation approach requires more robust rubric development through the establishment and definition of more specific assessment values for each metric in

Figure 1. Host Evaluation Criteria Hierarchy

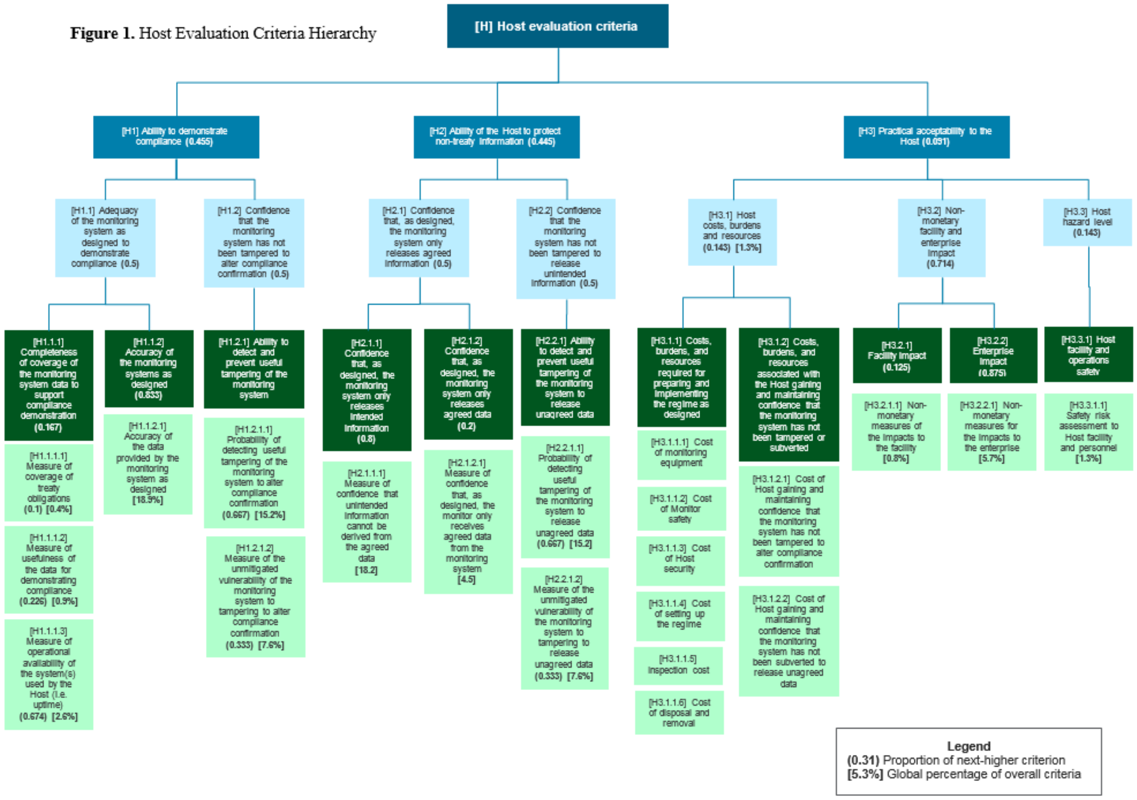
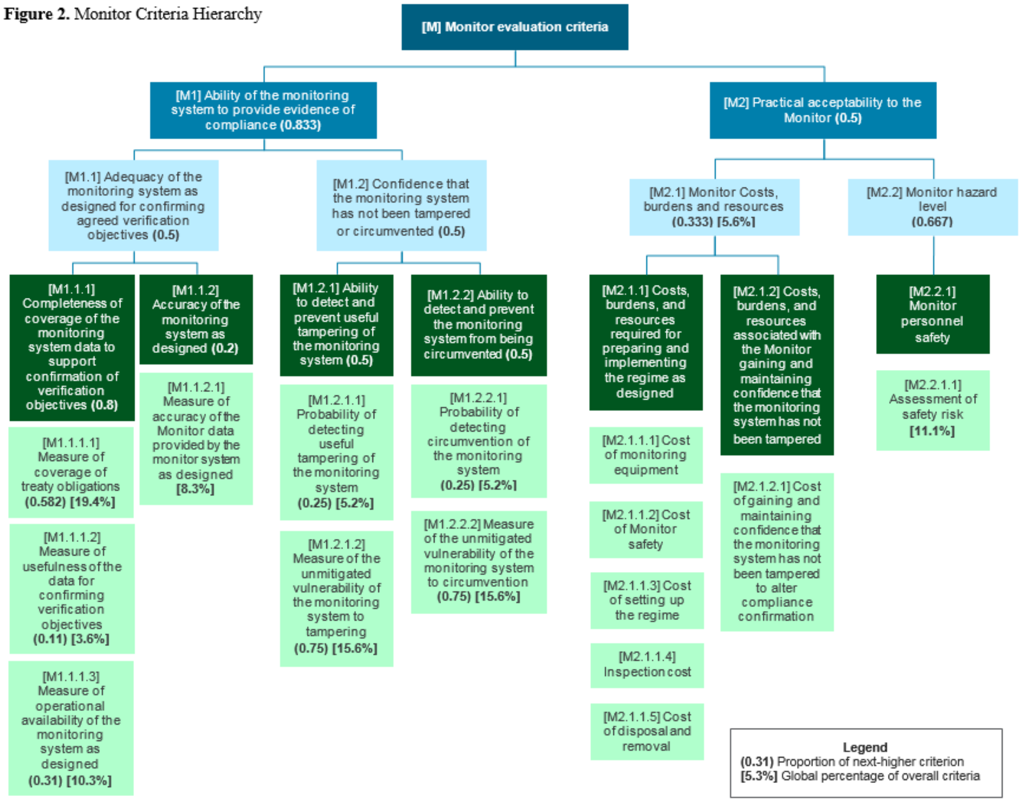

[H] Host evaluation criteria
[H1] Ability to demonstrate compliance (0.455)
[H2] Ability of the Host to protect non-treaty information (0.445)
[H3] Practical acceptability to the Host (0.091)
[H1.1] Adequacy of the monitoring system as designed to demonstrate compliance (0.5)
[H1.2] Confidence that the monitoring system has not been tampered to alter compliance confirmation (0.5)
[H2.1] Confidence that, as designed, the monitoring system only releases agreed information (0.5)
[H2.2] Confidence that the monitoring system has not been tampered to release unintended information (0.5)
[H3.1] Host costs, burdens and resources (0.143) [1.3%]
[H3.2] Non-monetary facility and enterprise impact (0.714)
[H3.3] Host hazard level (0.143)
[H1.1.1] Completeness of coverage of the monitoring system data to support compliance demonstration (0.167)
[H1.1.1.1] Measure of coverage of treaty obligations (0.1) [0.4%]
[H1.1.1.2] Measure of usefulness of the data for demonstrating compliance (0.226) [0.9%]
[H1.1.1.3] Measure of operational availability of the system(s) used by the Host (i.e. uptime) (0.674) [2.6%]
[H1.1.2] Accuracy of the monitoring systems as designed (0.833)
[H1.1.2.1] Accuracy of the data provided by the monitoring system as designed [18.9%]
[H1.2.1] Ability to detect and prevent useful tampering of the monitoring system
[H1.2.1.1] Probability of detecting useful tampering of the monitoring system to alter compliance confirmation (0.667) [15.2%]
[H1.2.1.2] Measure of the unmitigated vulnerability of the monitoring system to tampering to alter compliance confirmation (0.333) [7.6%]
[H2.1.1] Confidence that, as designed, the monitoring system only releases intended information (0.8)
[H2.1.1.1] Measure of confidence that unintended information cannot be derived from the agreed data [18.2]
[H2.1.2] Confidence that, as designed, the monitoring system only releases agreed data (0.2)
[H2.1.2.1] Measure of confidence that, as designed, the monitor only receives agreed data from the monitoring system [4.5]
[H2.2.1] Ability to detect and prevent useful tampering of the monitoring system to release unaqreed data
[H2.2.1.1] Probability of detecting useful tampering of the monitoring system to release unagreed data (0.667) [15.2]
[H2.2.1.2] Measure of the unmitigated vulnerability of the monitoring system to tampering to release unagreed data (0.333) [7.6%]
[H3.1.1] Costs, burdens, and resources required for preparing and implementing the regime as designed
[H3.1.1.1] Cost of monitoring equipment
[H3.1.1.2] Cost of Monitor safety
[H3.1.1.3] Cost of Host security
[H3.1.1.4] Cost of setting up the regime
[H3.1.1.5] Inspection cost
[H3.1.1.6] Cost of disposal and removal
[H3.1.2] Costs, burdens, and resources associated with the Host gaining and maintaining confidence that the monitoring system has not been tampered or subverted
[H3.1.2.1] Cost of Host gaining and maintaining confidence that the monitoring system has not been tampered to alter compliance confirmation
[H3.1.2.2] Cost of Host gaining and maintaining confidence that the monitoring system has not been subverted to release unagreed data
[H3.2.1] Facility Impact (0.125)
[H3.2.1.1] Non-monetary measures of the impacts to the facility [0.8%]
[H3.2.2] Enterprise Impact (0.875)
[H3.2.2.1] Non-monetary measures for the impacts to the enterprise [5.7%]
[H3.3.1] Host facility and operations safety
[H3.3.1.1] Safety risk assessment to Host facility and personnel [1.3%]
Legend
(0.31) Proportion of next-higher criterion
[5.3%] Global percentage of overall criteria

Figure 2. Monitor Criteria Hierarchy


[M] Monitor evaluation criteria
[M1] Ability of the monitoring system to provide evidence of compliance (0.833)
[M2] Practical acceptability to the Monitor (0.5)
[M1.1] Adequacy of the monitoring system as designed for confirming agreed verification objectives (0.5)
[M1.2] Confidence that the monitoring system has not been tampered or circumvented (0.5)
[M2.1] Monitor Costs, burdens and resources (0.333) [5.6%]
[M2.2] Monitor hazard level (0.667)
[M1.1.1] Completeness of coverage of the monitoring system data to support confirmation of verification objectives (0.8)
[M1.1.1.1] Measure of coverage of treaty obligations (0.582) [19.4%]
[M1.1.1.2] Measure of usefulness of the data for confirming verification objectives (0.11) [3.6%]
[M1.1.1.3] Measure of operational availability of the monitoring system as designed (0.31) [10.3%]
[M1.1.2] Accuracy of the monitoring system as designed (0.2)
[M1.1.2.1] Measure of accuracy of the Monitor data provided by the monitor system as designed [8.3%]
[M1.2.1] Ability to detect and prevent useful tampering of the monitoring system (0.5)
[M1.2.1.1] Probability of detecting useful tampering of the monitoring system (0.25) [5.2%]
[M1.2.1.2] Measure of the unmitigated vulnerability of the monitoring system to tampering (0.75) [15.6%]
[M1.2.2] Ability to detect and prevent the monitoring system from being circumvented (0.5)
[M1.2.2.1] Probability of detecting circumvention of the monitoring system (0.25) [5.2%]
[M1.2.2.2] Measure of the unmitigated vulnerability of the monitoring system to circumvention (0.75) [15.6%]
[M2.1.1] Costs, burdens, and resources required for preparing and implementing the regime as designed
[M2.1.1.1] Cost of monitoring equipment
[M2.1.1.2] Cost of Monitor safety
[M2.1.1.3] Cost of setting up the regime
[M2.1.1.4] Inspection cost
[M2.1.1.5] Cost of disposal and removal
[M2.1.2] Costs, burdens, and resources associated with the Monitor gaining and maintaining confidence that the monitoring system has not been tampered
[M2.1.2.1] Cost of gaining and maintaining confidence that the monitoring system has not been tampered to alter compliance confirmation
[M2.2.1] Monitor personnel safety
[M2.2.1.1] Assessment of safety risk [11.1%]
Legend
(0.31) Proportion of next-higher criterion
[5.3%] Global percentage of overall criteria

advance of the evaluation step. The development of these rubrics is not trivial. In fact, it could require additional research efforts to define specific assessment values for particular metrics in a way that enables meaningful differentiation between different monitoring systems. Additionally, the importance of monitoring system performance with respect to a particular metric is not necessarily absolute or constant. The importance of many of the criteria depends on many factors such as the relationship between the Host and Monitor countries, political will, and geopolitical stability. These factors would also need to be captured and used in rubric development.

According to an absolute approach, the performance of the monitoring system in each metric is measured according to a rubric or other grading scheme. Each metric may use a different rubric, but each rubric is normalized with the others so they can be combined to determine an overall score. The absolute approach produces a benchmarked score for a monitoring system that stands alone. However, the usefulness and meaningfulness of the absolute approach is limited by the specificity and accuracy of the normalized rubric used to assess the metrics. These rubrics help define and answer questions like, "how good is good enough" and "how do performance improvements within a metric translate to system-level gains?"

One option considered for the evaluation was the use of simple, 3-level rubric to evaluate the monitoring regimes: low, moderate, and high. While the effort to develop a simple rubric would be nearly negligible, the rubric lacked the specificity and accuracy needed to capture how small and large differences in metric performance impacted the overall performance of the monitoring system and the overall score. Lacking this, the evaluation did not provide accurate differentiation between each candidate monitoring system and was sensitive to subjective interpretation.

In cases where most of the data used to assess options is qualitative or unbenchmarked, a comparative evaluation approach such the Analytical Hierarchy Process (AHP) provides an accurate and effective approach to evaluate and compare options. This method does not require a rubric beyond the AHP scoring system. The resulting scores are not benchmarked in the same way as from an evaluation that uses an absolute approach. Because AHP scores the performance of each system relative to the competing alternatives, the scores are only meaningful within the set of alternatives that are compared. Furthermore, comparative approaches – particularly, the AHP – facilitate the incorporation of expert judgement in the scoring system. This is an important consideration as expert judgement will likely be a key input in scoring candidate monitoring systems, even as advancements are made in defining and measuring system performance.

FINDINGS

Informed by the challenges described above, the project team chose the AHP to evaluate and compare monitoring systems. According to this approach, the performance of the monitoring system in each metric is evaluated against the performance of other candidate monitoring systems. This method does not require a rubric beyond the AHP scoring system. The resulting scores are not benchmarked in the same way as the scores resulting from an evaluation that uses an absolute approach. Because AHP scores the performance of each alternative monitoring system relative to the competing alternatives, the score are only meaningful within the set of alternatives that are compared. This is one of the major limitations of the comparative approach: the results do not stand alone and scores for each monitoring system in the comparison group do not hold up when considered separately.

In determining which approach to use when evaluating monitoring systems, this effort also considered the intended use case to which this approach would be applied. The primary requirement of this use case is to provide logic and consistency for a more concrete evaluation of monitoring systems. Accordingly, this tool would be used primarily within the United States government agencies responsible for nuclear arms control treaty negotiation and verification to discuss options and develop a negotiating position rather than use during the treaty negotiation process with a treaty partner. Given this premise, this effort appreciated the flexibility of the AHP to adapt to changing political priorities and circumstances. These changes could be incorporated easily into the scoring process by applying the AHP to determine metric scores, which would likely take about 2 – 8 hours. Given the desire for a tool that enables a systematic, consistent comparison, the project team also considered the maturity of the elements required to fully develop and implement a system of evaluation using each approach. While it could require additional significant research efforts to fully develop a system of evaluation using the absolute approach, such development is not required for the comparative approach. Essentially, all that is required to implement a system of evaluation using the comparative approach is the development of a tool and documentation to support it.

Our recommendation to use a comparative approach is tied closely to the conditions of the expressed use case; there are conditions for which the absolute approach would be more appropriate. For example, if approaches are developed to quantitatively measure and compare the performance of a monitoring system a standard set of metrics, it may become appropriate to use the absolute approach. In that case, the absolute approach would provide a more rigorous method to compare monitoring systems against benchmarked assessment rubrics rather than against each other. Even still, these absolute scores may only support meaningful comparison within a single set of treaty considerations and assumptions (e.g. START). Similarly, because the comparative approach can incorporate quantitative measures to assess metrics (as is currently done with monetary costs), this approach could also benefit from the development of approaches to quantitatively measure the performance of a monitoring system in a certain metric. Using the comparative approach, these improvements could be incorporated one at a time as they become available and would not require the additional development of normalized rubrics.

LIMITATIONS

A driving principle in this reexamination is the desire to realize a set of criteria that are mutually exclusive and collectively exhaustive (MECE) [1]. Mutually exclusive criteria have no overlap and thus the criteria are not double counted. Collectively exhaustive criteria span the complete set of considerations that the Host and Monitor care about with respect to a monitoring system. While it is not possible to ensure perfectly MECE criteria, we have attempted to significantly improve those attributes of the criteria in this effort. Since many monitoring system factors can only be measured subjectively at the present time, it is difficult to ensure mutual exclusivity of the criteria. The system performance is based not only on technical data but also on the perceptions of each party, which may differ from one set of individuals to another. The team therefore must assume the perspective of each party in the absence of an actual treaty negotiation with identified parties and this will necessarily lead to some errors in judgement. In addition, there are likely to be disagreements about whether the criteria are mutually exclusive given their subjective nature. The team is not aiming to create a perfect tool; rather to create a useful one.

As formulated, the evaluation criteria are applied against the system as a whole (as applicable), rather than technology by technology or procedure by procedure. The criteria may potentially be applied at a lower level than the system, but that use of the criteria has not yet been analyzed. While this project was driven by the evaluation of CoC monitoring systems, the approach the project team followed to develop the evaluation criteria and roll-up methodology could be followed to develop criteria to evaluate other types of monitoring systems as well.

CONCLUSION AND NEXT STEPS

At the conclusion of this project, the project team successfully produced a set of criteria to help enable the systematic, repeatable evaluation of CoC monitoring systems. Additionally, the project team outlined a process to evaluate candidate monitoring systems against the new criteria and produce an overall score for the system. The team stresses that validation of this approach, for example through assessment of various use cases, is an important next step to enable completion of this research and, eventually, potential adoption of these criteria and the assessment approach.

Acknowledgements: Funding for this work, provided by the National Nuclear Security Administration's Office of Nuclear Verification, is gratefully acknowledged. The author acknowledges the team of expert who supported this research, including Matthew MacDougall (Pacific Northwest National Laboratory), Jay Brotz (Sandia National Laboratories), Danielle Hauck (Los Alamos National Laboratory), Daniel Krementz (Savannah River National Laboratory), Gary Cockrell (Pantex Plant), Gary Weeks (Crimson Sky Consulting).

REFERENCES

[1] Spencer, T. MECE Framework. http://www.spencertom.com/2013/01/30/mece-framework, 1/10/2018.