Defence Research and Development Canada
Recherche et développement pour la défense Canada

DEFENCE R&D DÉFENSE

INCOMMANDS TDP: Measures of performance for the threat evaluation capability

A. Frini
A. Benaskeur
DRDC Valcartier

Defence R&D Canada – Valcartier
Technical Memorandum
DRDC Valcartier TM 2009-240
December 2009

Canada

INCOMMANDS TDP: Measures of performance for the threat evaluation capability

Anissa Frini
Abderrezak Benaskeur
DRDC Valcartier

Defence R&D Canada – Valcartier

Technical Memorandum
DRDC Valcartier TM 2009-240
December 2009

Principal Author

Original signed by Anissa Frini

Anissa Frini

Defence Scientist

Approved by

Original signed by Stéphane Paradis

Stéphane Paradis

Head I2 Section

Approved for release by

Original signed by Christian Carrier

Christian Carrier

Chief Scientist

Abstract

In this document, we are concerned with the Threat Evaluation (TE) capability, developed for the INCOMMANDS TDP to support the operator involved in the shipboard TE process. In particular, we focus on evaluating the recommendations provided by this capability as well as evaluating the final TE decisions made by the operator when using this capability. A conceptual model explaining how these evaluations will be done is proposed. First, we propose a set of measures of performance (MoPs) for the assessment of the quality of the recommendations provided by the TE capability. These measures are: measures of accuracy and agreement, measures of timeliness, sensitivity analysis, input degradation analysis and anytime behaviour analysis. Second, we consider that the assessment of the level of cognitive activity that the human operator is experiencing is important. This evaluation will assess if higher quality of decision making is reached while the operator is overwhelmed or not. To do so, we propose the use of the following human factors metrics: situation awareness, workload and task relevant behaviour. Third, the quality of the TE decisions made by the operator (when using the capability) will be measured using complementary MoPs such as accuracy, timeliness and consistency. When analyzed together, these measures provide converging evidence that decision making quality has either increased or decreased.

Résumé

Dans ce document, nous nous intéressons à la nouvelle capacité d'évaluation des menaces développée dans le cadre du projet de démonstration technologique INCOMMANDS pour aider l'opérateur dans le processus d'évaluation des menaces. En particulier, nous nous concentrons sur l'évaluation des recommandations fournies par cette capacité ainsi que l'évaluation des décisions prises par l'opérateur en utilisant cette capacité. Un modèle conceptuel expliquant comment ces évaluations sont faites est proposé. Premièrement, nous proposons un ensemble de mesures de performance pour l'évaluation de la qualité des recommandations fournies par la capacité. Ces mesures sont : les mesures d'exactitude et d'accord, les mesures relatives aux temps d'exécution l'analyse de sensibilité, l'analyse de dégradation de la qualité des intrants et l'analyse du comportement "anytime" de l'algorithme. Deuxièmement, nous considérons que l'évaluation du niveau d'activité cognitive de l'opérateur est importante. Elle permettra d'évaluer si une meilleure qualité de la prise de décision est atteinte quand l'opérateur présente un niveau d'activité cognitif très élevé ou non. Ainsi, nous proposons l'utilisation des mesures de performance suivantes: éveil situationnel, charge de travail et comportements appropriés avec la tâche. Troisièmement, la qualité des décisions d'évaluation des menaces prises par l'opérateur sera mesurée en utilisant des mesures de performance complémentaires les unes aux autres telles que les mesures d'exactitude, les mesures relatives aux temps d'exécution et la consistance. Lorsque analysées ensemble, ces mesures multiples fournissent des preuves convergentes quant à l'amélioration ou non de la qualité de la prise de décision.

This page intentionally left blank.

Executive summary

INCOMMANDS TDP: Measures of performance for the threat evaluation capability:

A. Frini; A. Benaskeur; DRDC Valcartier TM 2009-240; Defence R&D Canada – Valcartier; December 2009.

Background: The Canadian HALIFAX Class ships are increasingly required to conduct operations in a littoral environment, which provides a new set of challenges in command and control (C2) for the Canadian Navy. Currently, the C2 functions are performed by operators in the operations room via a series of cognitive processes. While the operators are adept at performing these functions for a single threat, their ability to effectively achieve similar results for multi-threats is hampered. In addition, most current C2 systems have not been designed from a decision-centered perspective. The INCOMMANDS TDP was initiated with the intent to remedy this situation. It aims at developing and demonstrating advanced threat evaluation and combat power management decision support concepts for the command team of the Halifax Class Frigate in order *to improve the overall decision-making effectiveness*. This project aims at building the capability and at showing that it yields measurable performance benefits over the current system. Therefore, it is crucial to have a method which evaluates the quality of the decisions made by an operator when using the capability.

Results: In this document, we are concerned in particular with evaluating the recommendations provided by the threat evaluation (TE) capability developed for the INCOMMANDS project as well as evaluating the TE decisions made by the operator when using this capability. A conceptual model explaining how these evaluations will be done is proposed. The main results are:

1. A set of measures of performance (MoPs) for the assessment of the quality of the recommendations provided by the TE capability.

   Evaluating the quality of the recommendations given by the TE capability is crucial. In fact in operations room, the operator receives these recommendations via the functional displays, and after interpretation, he makes the final decisions. Thus, the quality of the recommendations must be evaluated since they have a great impact on the quality of TE decisions. Multiple measures of performance are proposed and formulated. In particular, measures of accuracy and measures of agreement are formulated to evaluate respectively the response of threat classification and ranking algorithms. Measures of timeliness are also considered as a complementary measure to accuracy. In addition to these MoPs, sensitivity analysis is considered to evaluate if small changes in the values of inputs have an impact on the results of the algorithm. Input degradation analysis is proposed to evaluate how the response quality of the algorithms behaves as function of decreasing input quality. Finally, anytime behaviour analysis is considered to evaluate the behaviour of the algorithms toward the runtime issues.

2. A set of human-factors metrics.

   We consider that the assessment of the level of cognitive activity that the human operator is experiencing is important. This evaluation will assess if higher quality of decision making is reached while the operator is overwhelmed or not. To do so, we propose the use of the following metrics: situation awareness, workload and task relevant behaviour.

3. A set of measures of performance for the assessment of the quality of TE decisions made by the operator when using the TE capability.

   The quality of the TE decisions made by the operator will be measured using multiple measures of performance such as accuracy, timeliness and consistency. When analyzed together, use of multiple measures provides converging evidence that decision making quality has either increased or decreased. To explain improvement or decrease of the decision making quality, evaluation of human-factors and evaluation of the performance of the TE capability presented earlier are useful.

Significance: The conceptual model as well as the measures of performance proposed in this work was thought originally in the spirit of evaluating the TE capability within the INCOMMANDS TDP. However, the results presented in this work are widely applicable for other similar problems. First, the conceptual model could be used in any other context where a capability supports a human operator in his decision-making process. The idea behind is to evaluate simultaneously the quality of the decisions made by the operator, the operator's performance and the quality of the recommendations provided by the capability. Second, measures of accuracy proposed to evaluate the quality of the recommendations provided by the capability remain appropriate for other capabilities that consist of classification of items inside categories (in particular the three-class classification problems). Also, measures of agreement for ranking algorithms remain suitable to compare two rankings of items in any other context. As well, sensitivity analysis, input degradation analysis and anytime behaviour analysis are sufficiently general and could be used in other situations.

Future plans: Measures of performance proposed in this work for the INCOMMANDS TDP could be used in future maritime projects as the Coalition Maritime Missile Defence (CMMD) TDP. Also, it could be used to evaluate other capabilities based on classification of items and for which the human operator intervene in the final decision (ex: object recognition and identification process).

Sommaire

INCOMMANDS TDP: Measures of performance for the threat evaluation capability:

A. Frini; A. Benaskeur; DRDC Valcartier TM 2009-240; R & D pour la défense Canada – Valcartier; Décembre 2009.

Introduction ou contexte: Les navires de la Classe Halifax sont de plus en plus appelés à mener des opérations dans un environnement littoral, qui fournit de nouveaux défis en commandement et contrôle (C2) pour les forces navales canadiennes. Actuellement, les fonctions C2 sont exécutées par des opérateurs dans les salles d'opération faisant appel à une série de processus cognitifs. Alors que l'opérateur est expert dans ces fonctions lorsqu'il s'agit d'une seule menace, ses capacités à réaliser la même performance pour plusieurs menaces sont entravées. De plus, la plupart des systèmes C2 ne sont pas conçus par rapport à une perspective centrée sur la décision. Le projet de démonstration technologique INCOMMANDS a été entrepris avec l'intention de remédier à cette situation. Il a pour objectif de développer et démontrer des concepts d'aide à la décision de commandement pour les processus d'évaluation des menaces et d'allocation des ressources dans des situations de guerre navale aérienne et de surface, afin d'*améliorer l'efficacité globale de la prise de décision*. Ce projet a pour but de concevoir le système et de démontrer que ce nouveau système réalise de meilleures performances comparé au système actuel. D'où, il est crucial d'avoir une méthode qui évalue la qualité de la décision prise par l'opérateur en utilisant la nouvelle capacité.

Résultats: Dans ce document, nous nous intéressons en particulier à l'évaluation des recommandations fournies par la capacité d'évaluation des menaces développée pour le projet INCOMMANDS de même que l'évaluation des décisions prises par l'opérateur en utilisant cette capacité. Un modèle conceptuel expliquant comment ces évaluations sont faites est proposé. Les principaux résultats sont :

1. Des mesures de performance de la qualité des recommandations fournies par la capacité d'évaluation des menaces.

   L'évaluation de la qualité des recommandations fournies par la capacité d'évaluation des menaces est cruciale. En effet dans la salle d'opération, l'opérateur reçoit des recommandations via les écrans d'affichage, et après interprétation, il prend les décisions finales. La qualité des recommandations doit être évaluée puisqu'elle a un impact important sur la qualité des décisions prises par l'opérateur. Plusieurs mesures de performance sont proposées. En particulier, les mesures d'exactitude et d'accord sont formulées pour évaluer respectivement les résultats des algorithmes de classification et d'ordonnancement des menaces. Les mesures relatives aux temps d'exécution sont aussi considérées comme des mesures complémentaires aux mesures d'exactitude. En plus de ces mesures de performance, l'analyse de sensibilité est examinée pour évaluer si des changements dans les valeurs des intrants auraient un impact sur les résultats des algorithmes. L'analyse de dégradation de l'input est envisagée pour évaluer comment l'algorithme se comporte lorsqu'une dégradation de la qualité de l'input est observée.

Finalement, l'analyse du comportement "anytime" de l'algorithme entre en compte pour évaluer le comportement des algorithmes avec le temps d'exécution.

2. Des mesures de facteurs humains.

   L'évaluation du niveau d'activité cognitif de l'opérateur est importante. Elle permet d'évaluer si une meilleure qualité de la prise de décision est atteinte quand l'opérateur présente un niveau d'activité cognitif très élevé ou non. Ainsi, nous proposons l'utilisation des mesures de performance suivantes: éveil situationnel, charge de travail et comportements appropriés avec la tâche.

3. Des mesures de performance de la qualité des décisions prises par l'opérateur en utilisant la capacité.

   La qualité des décisions d'évaluation des menaces prises par l'opérateur sera mesurée en utilisant des mesures de performance multiples telles que les mesures d'exactitude, les mesures relatives aux temps d'exécution et la consistance. Lorsque analysées ensemble, ces mesures multiples fournissent des preuves convergentes quant à l'amélioration ou non de la qualité de la prise de décision. Pour expliquer l'origine de l'amélioration ou de la détérioration de la qualité de la prise de décision, l'évaluation de la performance de l'opérateur et l'évaluation de la performance de la capacité sont utiles.

Importance: Le modèle conceptuel ainsi que les mesures de performance proposées dans ce travail ont été pensés à l'origine dans l'esprit d'évaluer la capacité d'évaluation des menaces dans le cadre du projet INCOMMANDS. Cependant, les résultats présentés dans ce document sont largement applicables pour d'autres catégories de problèmes similaires. Premièrement, le modèle conceptuel pourrait être utilisé dans n'importe quel contexte où une capacité aide un opérateur dans son processus de prise de décision. L'idée est d'évaluer la qualité des décisions prises par l'opérateur simultanément avec l'évaluation de sa performance et l'évaluation de la qualité des recommandations fournies par la capacité. Deuxièmement, les mesures d'exactitude proposées pour l'évaluation de la qualité des recommandations fournies par la capacité restent appropriées pour d'autres capacités fondées sur la classification d'éléments. Aussi, les mesures d'accord retenues pour les algorithmes de rangement peuvent être utilisées pour comparer deux rangements d'éléments dans n'importe quel autre contexte. De même, l'analyse de sensibilité, l'analyse de la dégradation de l'input et l'analyse du comportement "anytime" de l'algorithme sont suffisamment généraux et pourraient être utilisées dans d'autres situations.

Perspectives: Les mesures de performance proposées dans ce travail pour le projet INCOMMANDS pourraient être utilisées dans des projets maritimes futurs tels que le projet de démonstration technologique CMMD (*Coalition Maritime Missile Defence*). De même, ces mesures pourraient être utilisées pour évaluer d'autres capacités fondées sur la classification d'éléments et pour lesquelles l'opérateur humain intervient dans la prise de décision finale (ex : processus de reconnaissance et d'identification d'objets).

Table of contents

List of figures

List of tables

1 Introduction

The *Innovative Combat Management Decision Support Technology Demonstrator Project* (INCOMMANDS TDP) was initiated with the intent to develop and demonstrate advanced Above Water Warfare (AWW) Threat Evaluation (TE) and Combat Power Management (CPM) command decision support concepts for the command team of the Halifax Class Frigate in order to improve the overall decision-making effectiveness. This project aims at building a new capability and at showing that it yields measurable performance benefits over the current one. To achieve the required increase in net decision making effectiveness, the INCOMMANDS project assumes that the decision support systems and the human decision makers should form a joint-cognitive team by matching the human mental models with capability representations as algorithmic results, knowledge, and presentation models. Decision aids are in fact a crucial factor in augmenting cognition. They can make the problem task easier, and result in faster solution times, fewer steps to solve the problem, lower error rate, and quicker error recovery.

The INCOMMANDS TDP proposes these decision aids for both the TE and CPM processes. However, our focus in this document will be only on the threat evaluation process. Threat evaluation establishes the current intent and capability of non-friendly entities within the volume of interest based on a priori information (available intelligence, constraints), tactical picture, and data received from complementary sources in relation to the mission objectives. It is an ongoing process of determining if an entity intends and is able to inflict evil, injury, or damage to the defending forces and/or their interests, along with the prioritized ranking of such entities according to the level of threat they pose to the own-ship [1].

In this document, TE capability refers to the TE function developed within the Command Decision Support Capability (CDSC) of the INCOMMANDS project. It performs TE algorithms and presents TE recommendations to the operator via functional displays. The general objective of this work is to evaluate this capability as well as evaluate the TE decisions made by the operator when using this capability. More specifically, we try to:

1. Propose a set of measures of performance (MoPs) for the assessment of the quality of the recommendations provided by TE capability (algorithm's results).

2. Provide a set of human-factors metrics.

3. Propose a set of MoPs for the assessment of the quality of the TE decisions made by the operator when using the TE capability.

This technical memorandum is structured as follows. Chapter 2 presents the detect-to-engage process. In Chapter 3, the objectives and different components of the INCOMMANDS TDP project are presented. TE algorithms are specifically detailed. Chapter 4 is dedicated to the measures of performance for the TE capability. This chapter begins with a presentation of the conceptual model, which explains how to evaluate the TE decisions. Then, measures of performance for TE algorithms, human-factors measures as well as measures of the quality of TE decisions made by the operator (when using the TE capability) are proposed.

2 The detect-to-engage process

Command and Control (C2) is the means by which decision-makers synchronize military actions in time, space, and purpose to achieve unity of effort within a military force [2]. Advances in threat technology pose significant constraints and challenges to naval tactical detect-to-engage functions as they need to be processing high volume of data and information under time-critical conditions, for an increasing spectrum of difficult and diverse air, land, open-ocean and littoral scenarios.

The naval detect-to-engage process can be decomposed into a set of generally accepted functions that must be executed within some reasonable delays to ensure mission success [3]. A very high-level description of those functions, related to battle-space management, is given in Figure 1. This includes: picture compilation, threat evaluation, engageability assessment, and combat power management. The following sections give an overview of each function. A more detailed description of these processes is provided in [3], [4].

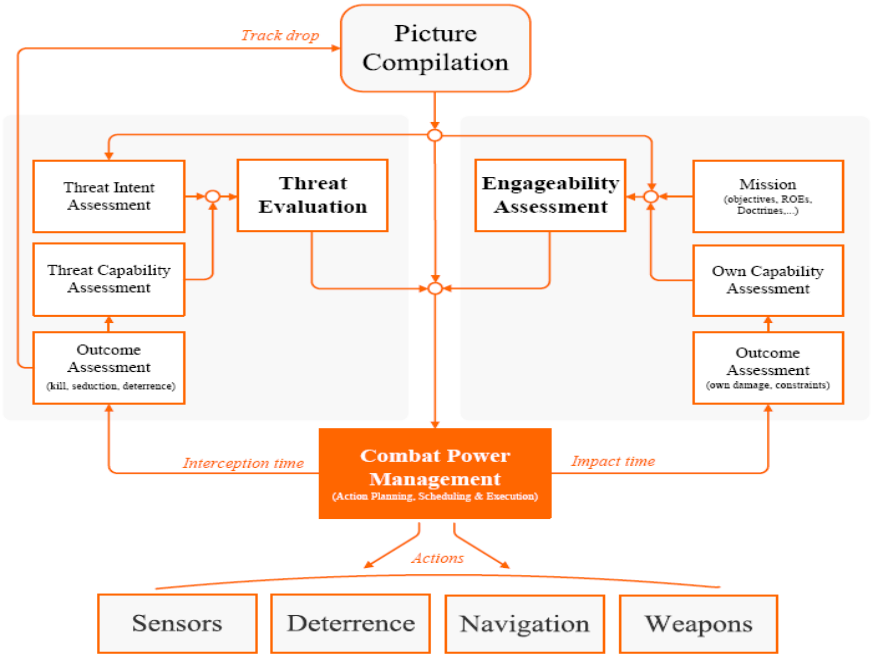

Figure 1: Naval C2 Process [3]

Figure 2 is related to the functional representation of Figure 1 and gives the state diagram for an object that enters the Volume Of Interest (VOI) of a defending force. The diagram shows the state transitions from detection to engagement and different underlying decisions. In this document, we are concerned specifically with the threat evaluation process. However, we will provide here a high-level description of all functions. The intent is to briefly describe the general conceptual framework in which to situate the TE process.

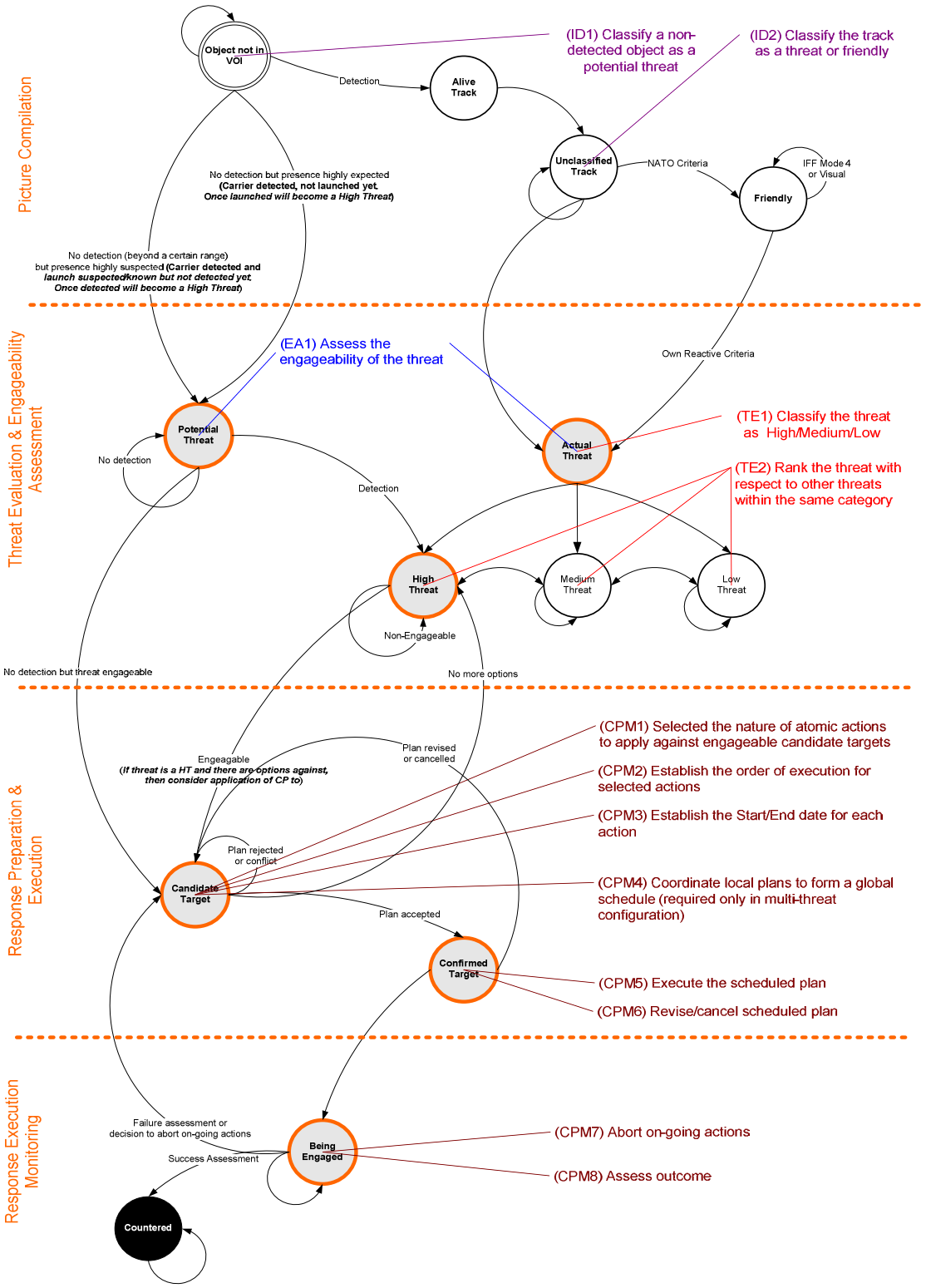

Figure 2: State diagram [3]

2.1 Picture compilation

In all maritime operations, ranging from peacetime to wartime, a fundamental requirement is the compilation of a plot of surface, air, and subsurface tracks. The process of all actions and activities aimed at compiling a plot is referred to as picture compilation. In maritime operations, picture compilation involves the following sub-processes [3], [4]:

1. Object detection: includes the employment of sensors in a volume of interest in order to determine the presence or absence of objects or object-related data;

2. Localization (or object tracking): includes the employment of sensors to keep track of the positional information and movements of an object;

3. Object recognition: includes the employment of sensors and other information to determine characteristics of an object. Comparing the collected characteristics against reference (or *a priori*) data can lead to correlation with a level of confidence; and

4. Object identification: includes the assignment of one of the six standard identities to a detected contact (hostile, suspect, unknown, neutral, assumed friend, friend).

2.2 Threat evaluation (TE)

Threat evaluation establishes the current intent and capability of non-friendly entities within the volume of interest based on a priori information, the tactical picture, available intelligence, constraints, and data received from complementary sources in relation to the mission objectives [1]. Threat evaluation involves the following sub-processes [1]:

1. Determining intent: this is a determination of whether the threat has hostile intent (or not) with respect to a reference point. Hostile intent translates into the will or determination of a threat to inflict harm or injury. Factors that may be considered in assessing the intent of a threat include: its combined speed/acceleration, heading and altitude with respect to a protected asset or defended area, the detection of emissions from its fire control radar, recognition of its preparing for operational readiness such as through refuelling, the estimation of its possible courses of action based on its pattern of movement and the events and activities in which it has participated (such as through its refusal to comply with repeated warnings issued by the defending forces for example), the recognition of its use of deception or tactics to evade being detected or tracked, its departure from an air corridor at any point other than designated exit points, and its response to own-ship manoeuvres [5], [6], [7].

2. Determining capability: this is a determination of the capability of a threat and involves evaluating both its opportunity and lethality. Opportunity involves evaluating if and when the threat has the ability to attack its target. Lethality is static information regarding the level of damage a threat can deliver if its combat power is applied against a target. The threat will only have the opportunity to deliver its lethality provided the following conditions are satisfied: the threat has sufficient energy to reach the target, the threat can detect and track the target, and physical obstructions do not impede access to the target.

2.3 Engageability assessment

Engageability assessment concerns the evaluation of own force's engagement options feasibility against the non-friendly objects within the VOI. This process is intended to help the CPM process (Section 2.4) by eliminating candidate solutions that violate one or more hard constraints. Engageability assessment involves the following sub-processes [3], [4]:

1. Mission Restraints: Impacts mission constraints on combat power (CP) deployment including rules of engagement (ROE); mission objectives; other warfare in progress or planned; and tactical doctrine based on a priori mission planning; and

2. Own Capability Assessment: This involves the estimation of the performance of combat power resources against individual threats, which requires the evaluation of the following two primary factors: determining the readiness of own combat power (availability and reliability) and predicting the performance of combat resources.

The output of this process is a list of combat power deployment options for each threat, with the associated degrees of freedom (e.g., time and range). This list of available options against each single threat is maintained with consideration of combined effects, synergy and usage constraints.

2.4 Combat power management (CPM)

In the context of naval C2, the CPM consists of the generation of a plan to use resources to counter selected threats. It also includes the monitoring and revision of combat plans under execution. Assessing the outcome of engagements is necessary to free up resources for use in subsequent combat power management plans. The CPM can be decomposed into three phases [3], [4]:

1. Response planning: ensures that one or more combat power resources are assigned to engage each target (i.e., atomic actions), including the assignment of supporting resources (e.g., sensors, communications). This involves assignment of both resources (i.e., pure allocation problem) and start and end times to activities (i.e., pure scheduling problem). A response (or engagement plan) is a coordinated (conflict-free) schedule (timeline) for the application of the selected combat power components.

2. Response Execution: involves executing in real-time the coordinated scheduled plan for the application of CP resources to counter targets within the current tactical situation.

3. Response Monitoring: is required since the responses are executed in a dynamic environment, subject to uncertainty, changing goals, and changing conditions. As such, the actual execution contexts will be different from the projected ones, i.e., the ones that motivated the construction of the original response. Monitoring is essential to help detect, identify and handle contingencies caused by uncertainty and the changing nature of the environment. Also involved is the evaluation of the outcome of the executed actions. This boils down to performing damage assessment (e.g., capability) of engaged target(s) and assessing the damage inflicted to own-assets by opponent forces.

3 INCOMMANDS TDP

This section provides an overview of the INCOMMANDS TDP and details the TE algorithms. Section 3.1 presents the objectives of the project. Then in section 3.2, the Command Decision Support (CDS) laboratory developed within the INCOMMANDS TDP is presented. Finally, Section 3.3 details the TE algorithms.

3.1 Objectives

The purpose of the INCOMMANDS TDP is to develop and demonstrate advanced AWW TE and CPM command decision support concepts for the command team of the Halifax Class Frigate in order *to improve the overall decision-making effectiveness* [8]. Specific objectives of the TDP are to:

1. Develop and demonstrate advanced AWW command decision support concepts in a manner that will assist the Halifax Class Modernization (HCM)/FELEX project define specifications for TE and CPM functions that are practicable for Canadian industry,

2. Elicit the Canadian Navy's cognitive/decision support and information requirements to perform single ship AWW command and control,

3. Develop a flexible and robust software architecture that enables the integration of heterogeneous algorithms and incremental enhancements,

4. Develop a knowledge-based framework that allows the efficient exploitation of *a priori* information and improves both human and automated TE/CPM functions,

5. Develop comprehensive evaluation methods and metrics (measures of performance and measures of effectiveness) that permit the empirical validation and assessment of new decision support systems and human decision-making effectiveness,

6. Develop an advanced naval C2 modeling and simulation capability that will be compatible with and of interest to the Canadian Forces Maritime Warfare Centre,

7. Explore multi-ship TE/CPM concepts in order to support the Canadian Navy's contribution to the international Battle Management Command, Control, Communications, Computers, and Intelligence project through a Task Group conceptual study, and

8. Demonstrate the developed command decision support concepts at sea by processing live data from a Halifax Class Frigate.

To support the purpose of the INCOMMANDS TDP, a Command Decision Support Laboratory (CDS Lab) has been developed in order to allow the evaluation of the decision support concepts. This includes both the development of the decision support concepts themselves (i.e., new operator displays), the development of the information processing component required to drive the displays (i.e., threat evaluation and combat power management algorithms), as well as the

development of a synthetic environment within which scenarios may be executed to stimulate the information processing component and the new displays. Section 3.2 provides an overview of the CDS laboratory and its components.

3.2 CDS laboratory overview

The CDS laboratory built in this project is composed of a Command Decision Center (CDC) and a *test bed* (*Figure 3*). The CDC is composed of Operator/Decision-Makers and computer-based Command Decision Support Capability (CDSC). The role of the operator is to make good and timely decisions based on the information conveyed by the CDSC.

The *testbed* environment (experimentation capability) is defined in this project as the set of capabilities that allow for a realistic stimulation and performance measurement of the system while achieving the decision-making process investigated in this project. More specifically, the testbed consists of hardware and software requirements that must be implemented to ensure that the experimental objectives can be met.

The CDS laboratory architecture is provided in Figure 4.

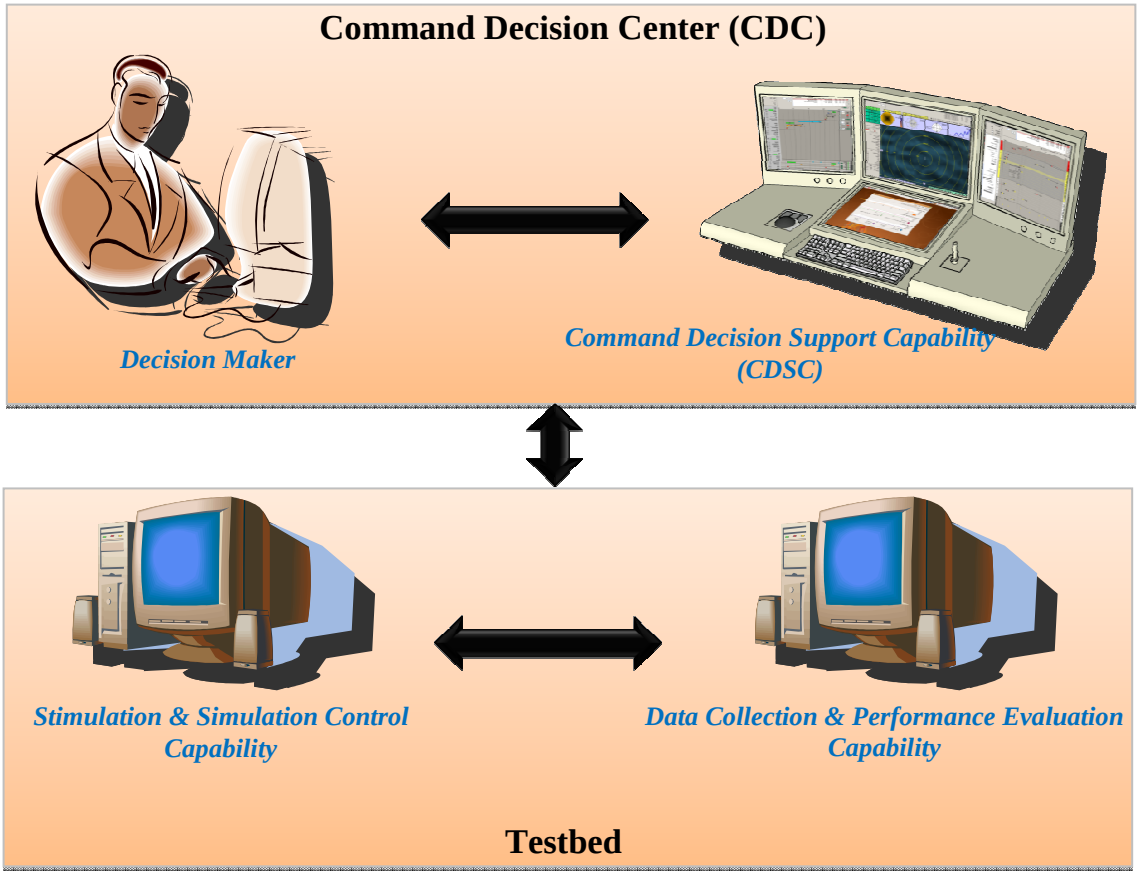

Figure 3: CDS Lab overview

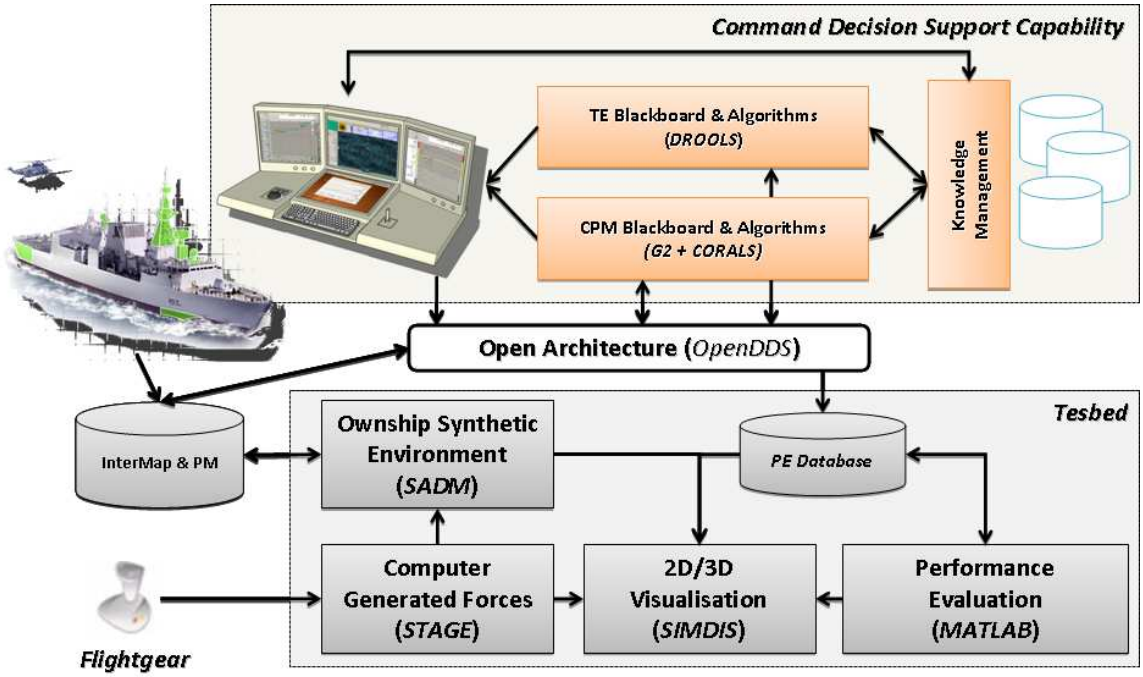

Figure 4: CDS Lab architecture

The INCOMMANDS CDSC

The INCOMMANDS CDSC is composed of:

1. A TE & CPM component, which consist of the algorithms related to threat evaluation and combat power management. These algorithms are used to provide a prioritized threat ranking as well as to generate and execute a combat power management plan for targets that are to be engaged.
2. A functional Operator Machine Interface (OMI) component which presents TE & CPM data to the operator and accepts operator (TE/CPM-related) inputs and queries.
3. A physical OMI component which presents a "conventional" Plan Position Indicator (PPI) like display that presents tracks and associated information.
4. A knowledge Management component which delivers a priori information to the TE/CPM component and to operators via the Functional OMI component.

TE Capability

Within the INCOMMANDS TDP, TE capability refers to the computer-based capability performing TE algorithms and presenting TE recommendations to the operator via functional OMI displays. More specifically, TE capability refers to the sub-set of the CDSC which is composed of:

i. A TE functional OMI component which visualizes the TE recommendations to the operator and interacts with him.

ii. A TE information processing component, which consists of TE algorithms;

Illustration

An illustration of a conceptual INCOMMANDS prototype comprising multiple single-role displays is provided in Figure 5. The console is comprised of four interdependent types of displays: TE functional view, CPM functional view, physical view, and general purpose. The complementary advantages provided by each display are envisioned to augment the operator's decision-making capabilities.

The TE functional display is comprised primarily of a 2-dimensional threat list. The list presents a relative ranking of threats based on their threat rating as calculated by the CDSC using the appropriate TE algorithm. To that end, the operator is able to monitor the complete threat picture as well as perform pre-emptive planning as required to avoid a potential engagement [4]. Figure 6 provides a depiction of the TE functional display. The CPM functional view is aimed at supporting the management and application of combat power through a series of decision aids. The CPM functional view consists of four predominant groupings: target-based view of engagement plans, resource-based view of engagement plans, engagement drill downs, and notifications [4]. Figure 7 provides a depiction of the CPM functional display.

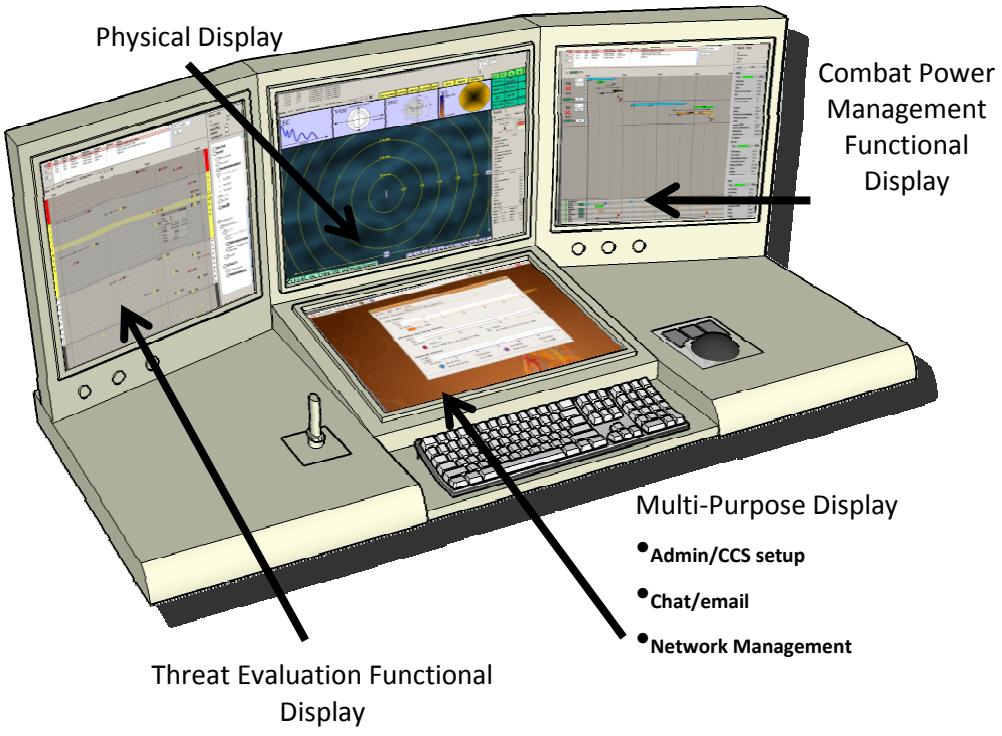

Figure 5: Conceptual console for single-role displays

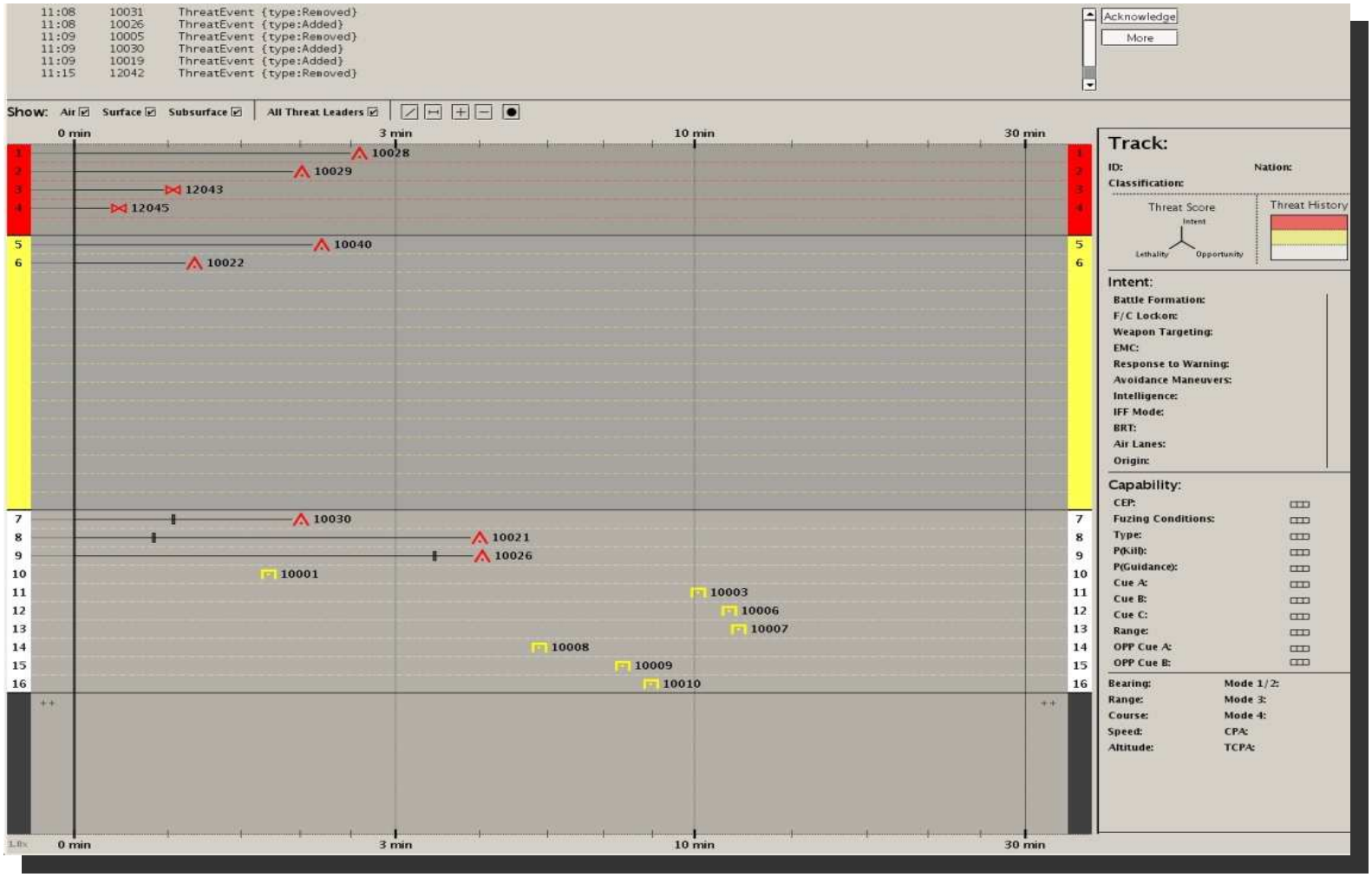

Figure 6: Threat evaluation functional view

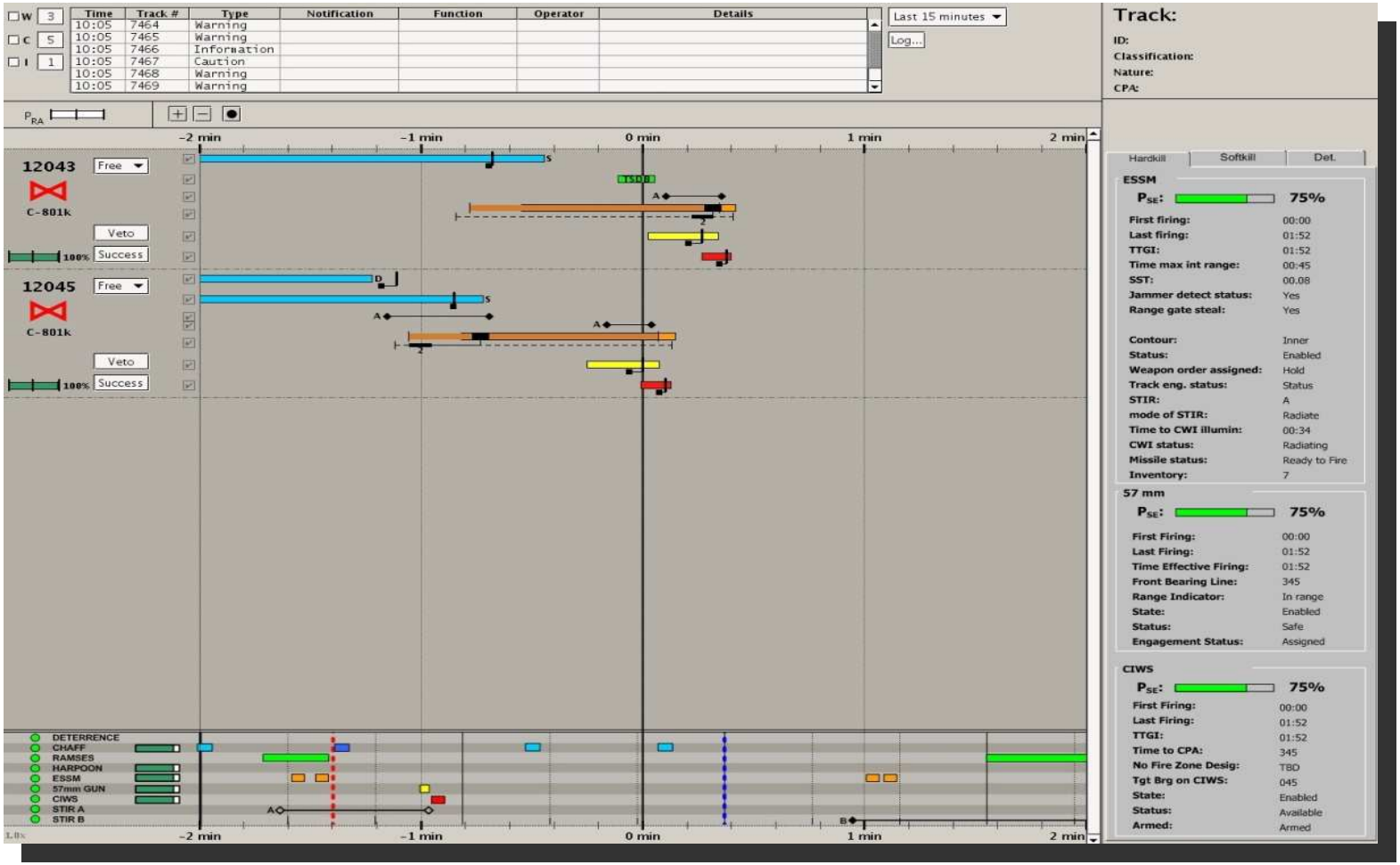

Figure 7: Combat power management functional view

3.3 Threat evaluation algorithms

In this section, we will be focusing on the TE algorithms. These algorithms consist of determining if a non-friendly entity within a certain volume of interest has capability and intends to inflict evil, injury, or damage to the defending forces and its interests, along with the ranking of such entities according to the level of threat they pose (Figure 8).

TE algorithms relies heavily on the use of the compiled tactical picture and the available contextual information such as the locations of vital assets and defended regions, attributes of platforms, weapon systems and surveillance systems, doctrine, intelligence reports, information about features of the terrain and the area of operations, knowledge of the opposing force's structure and the recent history of its behaviour in the operation area [1], [9].

3.3.1 Performing reactive test and VOI test

In support to TE algorithms, two tests are performed first: the reactive test and Volume Of Interest (VOI) test. For threats which do not pass the reactive test and pass the VOI test, TE algorithms will consist of the following three sub-functions [8]: “Determine hostile intent”, “Determine capability” and “Categorize and rank threats”.

Reactive test and VOI test are performed on tracks generated by the picture compilation process.

Reactive test: The reactive test consists of identifying - with a high degree of certainty - highly threatening tracks that have not to pass through the TE process. These tracks have specific properties (e.g. high speed, incoming, and low altitude) that let the track to be considered as a High Threat.

VOI test: The VOI test aims at limiting the set of system tracks that will be subjected to threat evaluation capability. The idea behind the VOI test is to avoid operator cognitive overload, by not dedicating mental and computational resources to tracks which do not pose an immediate threat [8].

Let us note that the TE capability allows the operator to specify both the reactive test criteria and the VOI criteria, subject to some constraints. Annex A.1 provides a specification of the default reactive criteria and the VOI criteria respectively.

3.3.2 Determining hostile intent

“Determining hostile intent” is a sub-function within the TE function which determines whether or not a threat is exhibiting hostile intent. All threats will be subjected to continual (over time) re-assessment of hostile intent, except highly threats that have passed the reactive test. The assessment of hostile intent involves a determination of specific intent-disclosing behaviours.

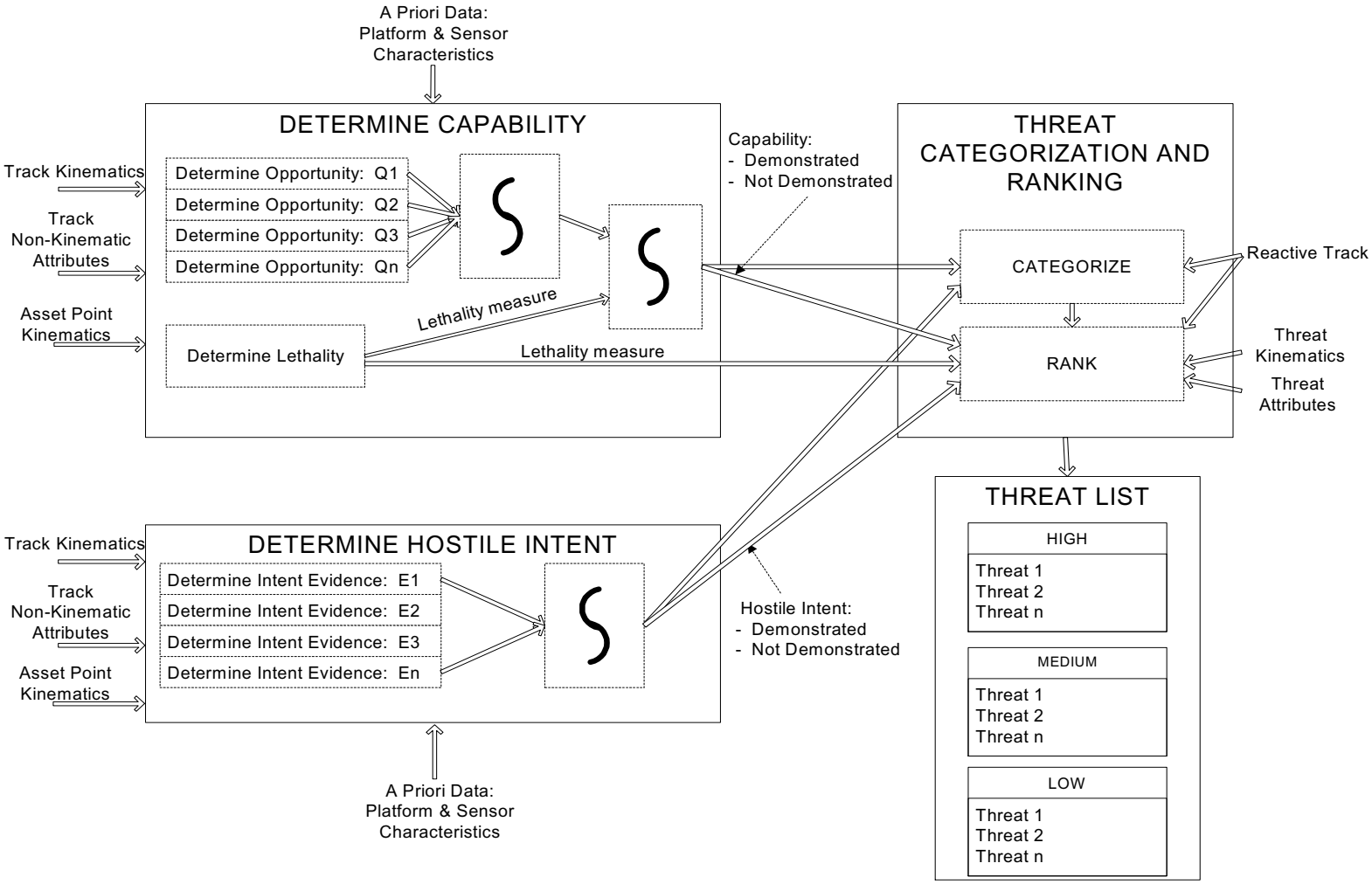

Figure 8: TE algorithms [8]

The sub-function is operator-configurable so that, during mission preparation, the operator is able to exercise some degree of control over the criteria that will be used to establish the presence of hostile intent. Annex A.2 provides the default criteria for the "determining hostile intent" sub-function considered in the INCOMMANDS CDSC. This flexibility is required since the assessment of threat intent depends on a number of factors that can vary from mission to mission. As part of the mission planning activity, the operator will specify a set of behaviours indicating hostile intent as well as a set of behaviours indicating non-hostile intent. Then, the operator will characterize the behaviours in terms of their sufficiency and necessity to the establishment of hostile intent[1]. A pseudo-code articulation of the "determining hostile intent" sub-function [8] is:

> IF [any combination of behaviours *sufficient* to establish hostile intent is present] AND IF [all behaviours *necessary* to establish intent are present] THEN <Hostile Intent is established> ELSE <Hostile Intent is not established>.

3.3.3 Determining capability

"Determining capability" is a sub-function within the TE function which determines whether or not a threat has the capability to harm an asset point. All threats will be subjected to continual (over time) re-assessment of their capability, except threats that have passed the reactive test. Assessment of threat capability involves the estimation of relational and non-relational properties of the threat [8]. Examples of non-relational properties include explosive payload, munitions inventory, and phase of operation (e.g. search phase, fire control phase, etc.). Examples of relational properties include threat stored energy sufficiency to reach the defended asset, terrain obstructions of threat flight to the defended asset, and presence of defended asset within the field of view of target sensor(s) [8].

The capability of a threat to harm the defended assets point is constituted by its "lethality" and its "opportunity to deliver the lethality to the defended asset and satisfy conditions for successful detonation" [8]. The determination of threat **opportunity** is multi-dimensional. The operator will have some degree of freedom to set the opportunity conditions (Annex A.3). If opportunity conditions are satisfied in all of its dimensions, the presence of opportunity is established. **Lethality** is an entirely intrinsic property of the threat, and more specifically a measure of the magnitude of its explosive payload.

The way that assessment of opportunity and assessment of lethality are combined in order to produce an overall assessment of capability is largely determined by the operator through the setting of configuration data. The range of possible outputs generated by this sub-function includes: i) threat has capability to harm the asset point and ii) threat does not have capability to harm the asset point.

[1] For example, a radio announcement from a threat indicating an intention to attack might be considered a sufficient but not necessary condition for concluding that the threat exhibits hostile intent. The presence of this behaviour establishes intent regardless of the presence or absence of other behaviours that suggest hostile intent. However, the absence of such behaviour does not rule out the establishment of hostile intent. By contrast, the behaviour of "heading straight for the reference point" (e.g. low CPA behaviour) might be chosen to be a behaviour that is necessary but not sufficient for establishing hostile intent. In that case, the behaviour must be present in order to establish hostile intent, but other behaviours could be present as well.

3.3.4 Threat classification and ranking

The threat classification algorithm is a rule-based anytime classifier. It classifies each threat in the corresponding category (high, medium, low) depending on a set of rules based on their intent, capability to deliver damage (lethality & opportunity) and other factors (e.g. reactive test results, threat kinematics, and threat non-kinematics attributes). Table 1 details the threat classification mechanism as implemented in the INCOMMANDS TDP.

Table 1: Threat classification mechanism

Threat Category	Hostile Intent	Capability
High	Pass Reactive Test	
	Yes	Yes
Medium	Unknown	Yes
	Yes	Unknown
	Yes	No
	No	Yes
Low	Unknown	Unknown
	No	Unknown
	Unknown	No
	No	No

This sub-function is operator-configurable so that, during mission preparation, the operator is able to modify the conditions that establish the category membership criteria for each threat (default criteria used in the CDSC to categorize threats are presented in Table 1). Let us note however that the operator cannot modify, for example, the number of threat categories.

The threat ranking algorithm consists of ranking threats inside each category (high, medium, low). Within each of the three threat categories, threats are ranked in descending order of lethality. In the event one or more threats (in the same category) have the same lethality value, such threats are ordered according to the track number - the lower the track number, the higher the ranking.

Although the classification and ranking algorithms implemented in the INCOMMANDS TE capability are quite simple, the MoPs that we will propose in Section 4.2 could be applicable for more complex threat classification and ranking algorithms since the MoPs don't depend on the way the algorithm works.

4 Measures of performance for TE capability

In this Chapter, we will start by presenting the evaluation methodology of the TE capability. Then, the details of the measures of performance are provided.

4.1 Evaluation methodology

The INCOMMANDS TE capability aims at developing TE command decision support concepts for the command team of the Halifax Class Frigate in order *to improve the overall TE decision-making effectiveness*. Therefore, it is crucial to have a method on which the quality of TE decision-making is evaluated. Such method will answer the following question "how to measure the quality of TE decisions provided by the INCOMMANDS system given that these decisions are made by an operator when using the TE capability.

Let first point out some terminology. The **TE system** refers to the combination of the TE capability and the operator. **TE capability** refers to a computer-based capability performing TE algorithms and presenting TE recommendations to the operator via functional OMI displays. **Operator** refers to the military person who has access to the functional and physical displays, receives recommendations from the TE capability and makes final TE decisions (classification and ranking of threats). **TE recommendations** refer to the output of the TE algorithms implemented in the capability and **TE decisions** refer to the decisions made by the operator when using the TE capability.

The evaluation methodology has the following objectives:

1. Provide MoPs for the TE algorithms implemented in the capability. These MoPs assess the quality of the recommendations given to the operator by the TE capability. This evaluation verifies if higher quality of the decision-making is associated or not to the operator. In other words, this evaluation aims at identifying the case where the quality of the decisions made by the operator is high even if the quality of the recommendation is not good (ex: not accurate, not timely, etc.). In such cases, the quality of the decision making will be operator-dependent. A more experienced operator will provide better results given that he is able to change the classification and/or ranking of threats in case he perceives errors in the recommendation.

2. Provide human-factors metrics in order to assess the level of cognitive activity that the human operator is experiencing when using the system. This evaluation verifies if higher quality of decision making is accompanied or not by an overwhelming of the operator. In fact, the decision aids, when ideally designed, should make the problem task easier for the operator and result in better situation awareness, lower workload, faster solution times, fewer steps to solve a problem, lower error rate, and quicker error recovery.

3. Provide MoPs for the TE decisions produced by the INCOMMANDS TE system (operator using the TE capability). These MoPs assess the quality of the TE decisions in terms of accuracy, timeliness and consistency of the decisions.

Figure 9 illustrates a conceptual model summarizing how to measure the quality of TE decisions provided by an operator when using the TE capability. This conceptual model assumes that the decision-making quality is explained by both the human factors and the performance of the TE capability (in term of algorithm's results).

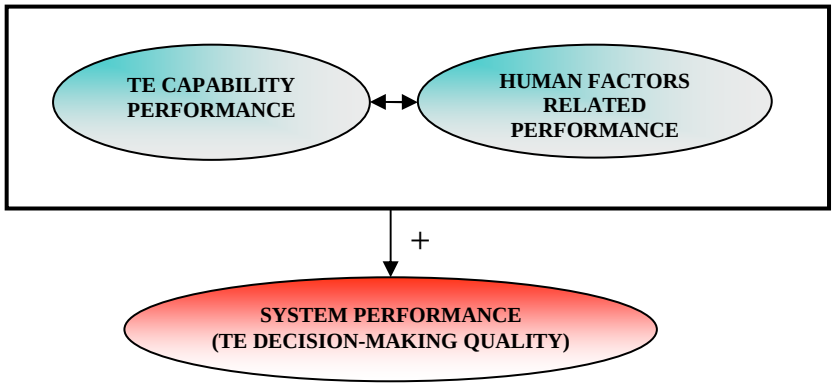

Figure 9: Conceptual model

The measurement of TE decision-making quality will be based on this model and a multi-dimensional approach will be considered. A series of measures that assess the quality of decision making is proposed in Figure 10.

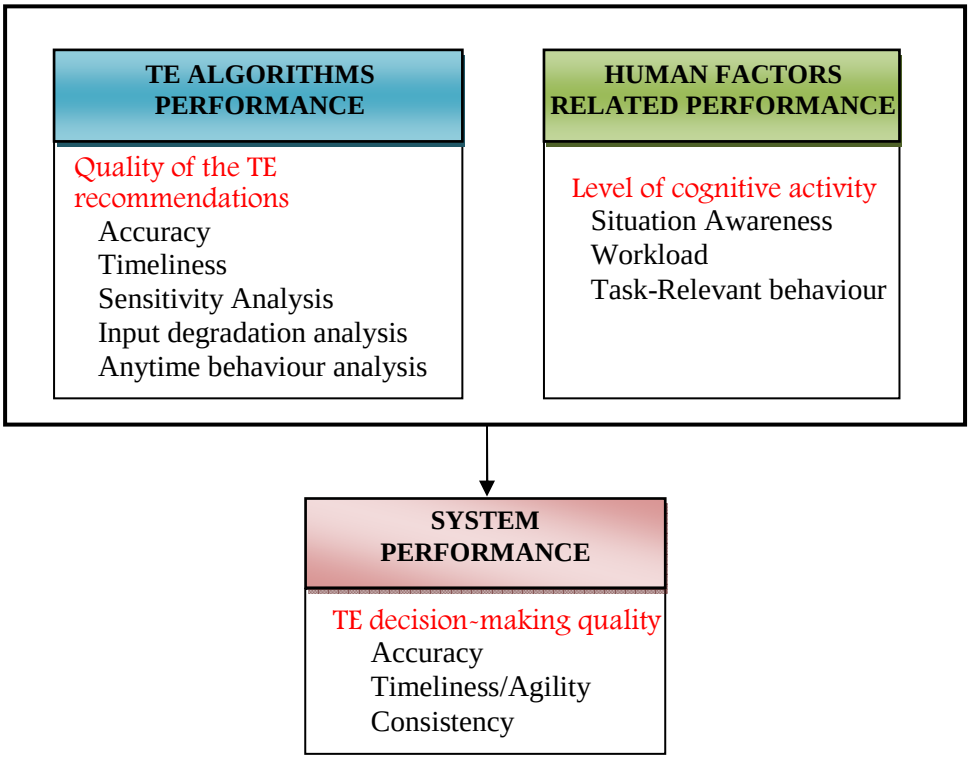

Figure 10: Measures of performance

First, the quality of the TE decisions made by the operator will be measured using multiple MoPs such as accuracy, timeliness and consistency. When analyzed together, use of multiple measures provides converging evidence that decision making quality has either improved or decreased.

However, to explain such improvement or decrease of decision making quality, we should evaluate both human-factors performance and TE capability performance. First, we consider that the assessment of the level of cognitive activity that the human operator will be experiencing is important. This evaluation will assess if higher quality of decision making is reached while the operator is overwhelmed or not. To do so, we propose the use of the following MoPs of the operator: *situation awareness, workload and task relevant behaviour*. Second, we consider that evaluating the quality of the recommendations given by the TE capability is crucial. In fact in operations room, the operator receives these recommendations via the functional displays, and after interpretation, he makes the final decisions. Thus, the quality of the recommendations must be evaluated since they have a great impact on the quality of TE decisions. Proposed measures of performance of the threat classification and ranking algorithms are: *accuracy measures, timeliness, sensitivity analysis, input degradation analysis, anytime behaviour analysis*.

The following sections will detail the measures presented here and constitute the core part of this document. First, MoPs for TE algorithms are defined and formulated. Then, human factors measures are provided. Finally, measures of TE decision-making quality are presented.

4.2 Measures of performance for TE algorithms

In this chapter, our focus is on how to measure the performance of threat classification and ranking algorithms.

Performance evaluation of the TE algorithms will be done in two steps. In the first step, we consider a scenario with a large number of threats, which are already classified and ranked by one (or a group of) Subject Matter Expert (SME). These threats are classified and ranked by the algorithms and results are compared with the SME classification and ranking. Accuracy measures for threat classification algorithms and agreement measures for threat ranking algorithms (detailed below in Section 4.2.1 and Section 4.2.2) will be used to assess the accuracy of the results for the considered scenario. In addition, timeliness measures are important to consider because generally, runtime and quality of algorithm results are conflicting criteria (runtime is traded for quality of results). Then, sensitivity analysis, input degradation analysis and anytime behaviour analysis will be performed for this scenario. In the second step, we consider other scenarios (in different context) that have different levels of complexity. For each scenario, we compare again the classification and ranking of threats by the algorithms with the classification and ranking of the SME and we evaluate timeliness of the algorithm. Then we perform sensitivity analysis, input degradation analysis and anytime behaviour analysis. The advantage of considering different scenarios in different context is twofold. First, evaluating the algorithm in different context is crucial since the processing of data and the results could differ significantly with the context (e.g. the level of threat of the same object could be high in one context and much lower in another context). This context-based analysis aims at verifying that the quality of the output (captured with accuracy measures) remains good whatever the scenario and whatever the context. Second, considering different scenarios that have different level of complexity is important since it allows analyzing the algorithm's results over the complexity of the scenario.

4.2.1 Accuracy measures for threat classification algorithms

Literature review shows that several accuracy measures are proposed to assess the performance of the classification algorithms. These metrics were developed specifically to assess the performance of binary (two class) classification algorithms. The most-known metric is the misclassification rate, which is largely criticized [10] because of its insensitivity to the type of error (assuming equal misclassification cost) and its assumption of uniform distribution of classes. In addition, the misclassification rate is not very relevant if the prevalence of the data set is very low. The limitation of these accuracy measures could be overcome if the cost of misclassification is taken into consideration. However, these costs are difficult to determine [10],[11],[12]. An alternative way to measure the performance of the algorithms consists of using other metrics introduced earlier by [13]. These metrics are based on the confusion matrix defined in [11] as "a form of contingency table showing the differences between the true and predicted classes for a set of labelled examples" (see Annex B). Consequently, other metrics based on the confusion matrix, were suggested in the literature such as: sensitivity, specificity, false positive rate, false negative rate, positive and negative predictive power, etc. (see Table 11 in Annex B). Let us note that there are other interesting measures (Receiver Operating Characteristic (ROC) curve, the area under ROC curve) that are widely used when the classifier output is continuously distributed (objects are assigned to categories depending on a threshold value). However, these two measures could not be used in our study because the classifier output of the algorithm is discrete.

For threat classification algorithms of the INCOMMANDS TDP, we are concerned specifically with three-class classification algorithms. Thus, we propose to generalize the metrics based on the confusion matrix in case of binary classification (Table 11), to the case of three class classification.

The generalization of the confusion matrix to a three class context is given in Table 2.

Table 2: Confusion matrix for three-class classification

	Actual "High"	**Actual "Medium"**	**Actual "Low"**
Assessed "High"	T_H	$F_{H/M}$	$F_{H/L}$
Assessed "Medium"	$F_{M/H}$	T_M	$F_{M/L}$
Assessed "Low"	$F_{L/H}$	$F_{L/M}$	T_L

Where T_i is the number of threats correctly classified by the algorithm in category i ($i = H, M, L$ where H is "High", M is "Medium" and L is "Low") and $F_{i/j}$ is the number of threats actually in category j (for $j = H, M, L$) but incorrectly classified by the algorithm in category i. N is the total number of threats. Assessed i for $i = H, M, L$ represents the classification given by the algorithm concerning a threat and Actual i for $i = H, M, L$ represents the real classification of the threat usually given by Subject Matter Experts (SMEs).

The following sub-sections provide the accuracy measures resulting from the generalization we have done. We propose measures of overall misclassification rate, sensitivity of class i (% threats actually in class i that are correctly classified), algorithm errors of type 1 (errors in classification resulting in a jump of two categories), algorithm errors of type 2 (errors in classification resulting in a jump of one category), algorithm predictive power for class i (probability that a threat is actually in category i if the model classifies it in category i). In addition, the Kappa statistic, reported from the literature, is also presented as a measure of performance.

Overall misclassification rate

The overall misclassification rate ρ refers to the percentage of threats for which the assessed category by the algorithm is incorrect. The calculation of the overall misclassification rate is given by Equation (1). Misclassification rate varies in [0, 1]. A value of 1 is indicative of poor performance (high misclassification rate) whereas a value of 0 is indicative of good performance (low misclassification rate).

$$\rho = 1 - \frac{T_H + T_M + T_L}{N} \tag{1}$$

The same measure could be processed only for threats within a sufficient range to harm the ship. It will correspond to the percentage of threats, within a sufficient range to harm the ship, for which the assessed category is incorrect.

Sensitivity of class i

Sensitivity of class i for $i = H, M, L$ is the conditional probability that a threat actually in class i is classified correctly by the algorithm. Sensitivity varies in [0, 1]. A value of 1 is indicative of high performance whereas a value of 0 is indicative of poor performance. These measures are given below.

S^H is the sensitivity of class "High". It refers to the percentage of threats actually "High" that are correctly classified; as shown by Equation (2).

$$S^H = \frac{T_H}{T_H + F_{M/H} + F_{L/H}} \tag{2}$$

S^M is sensitivity of class "Medium". It refers to the percentage of threats actually "Medium" that are correctly classified; as shown by Equation (3).

$$S^M = \frac{T_M}{T_M + F_{H/M} + F_{L/M}} \tag{3}$$

S^L is sensitivity of class “Low”. It refers to the percentage of threats actually “Low” that are correctly classified; as shown by Equation (4).

$$S^L = \frac{T_L}{T_L + F_{H/L} + F_{M/L}} \tag{4}$$

Algorithm errors (Type 1)

The algorithm errors of type 1 refer to the misclassification resulting in a jump of two categories. In other words, the algorithm assesses the threat as “High” while it is in category “Low” and vice versa. These type of errors are very harmful for the own ship since a threat could be considered as low threatening while it is in the “High” category and needs to be engaged. In the other case, the threat could be considered by the algorithm as high threatening and so needs to be engaged while it is actually in the “Low” category. Algorithm errors of type 1 should be avoided in all circumstances. These measures are given below.

$\varepsilon_{L/H}$ is the percentage of threats actually “High” classified in the category “Low”, as calculated by Equation (5).

$$\varepsilon_{L/H} = \frac{F_{L/H}}{T_H + F_{M/H} + F_{L/H}} \tag{5}$$

$\varepsilon_{H/L}$ is the percentage of threats actually “Low” classified in the category “High”, as calculated by Equation (6).

$$\varepsilon_{H/L} = \frac{F_{H/L}}{T_L + F_{H/L} + F_{M/L}} \tag{6}$$

These MoPs vary in [0, 1]. A value of 1 is indicative of extremely poor performance whereas a value of 0 is indicative of high performance (no errors of type 1).

The same measure could be processed only for threats within a sufficient range to harm the ship. It corresponds to the percentage of threats - actually “Low” and positioned within a sufficient range to harm the ship- that are classified in category “High” and the percentage of threats -actually “High” and positioned within a sufficient range to harm the ship- that are classified in category “Low”.

Algorithm errors (Type 2)

The algorithm errors of type 2 refer to the misclassification resulting in a jump of one category. This type of error could imply disastrous consequences like the algorithm errors of type 1, especially when the assessed category is “High” while it is actually “Medium” and vice versa. In fact, a “Medium” threat might be engaged and a “High” threat might be ignored. However, the

consequences are less harmful if the assessed category is “Medium” while it is actually “Low” and vice versa. In this specific case, there will not be any error in engagement because only “High” threats are engageable. These measures are given below.

$\varepsilon_{M/H}$ is the percentage of threats actually “High” classified in the category “Medium”, as calculated by Equation (7).

$$\varepsilon_{M/H} = \frac{F_{M/H}}{T_H + F_{M/H} + F_{L/H}} \tag{7}$$

$\varepsilon_{H/M}$ is the percentage of threats actually “Medium” classified in category “High” as calculated by Equation (8).

$$\varepsilon_{H/M} = \frac{F_{H/M}}{T_M + F_{H/M} + F_{L/M}} \tag{8}$$

$\varepsilon_{M/L}$ is the percentage of threats actually “Low” classified in the category “Medium” as calculated by Equation (9).

$$\varepsilon_{M/L} = \frac{F_{M/L}}{T_L + F_{M/L} + F_{H/L}} \tag{9}$$

$\varepsilon_{L/M}$ is the percentage of threats actually “Medium” classified in the category “Low” as calculated by Equation (10).

$$\varepsilon_{L/M} = \frac{F_{L/M}}{T_M + F_{H/M} + F_{L/M}} \tag{10}$$

These MoPs vary in [0, 1]. A value of 1 is indicative of poor performance whereas a value of 0 is indicative of high performance (no errors of type 2).

Algorithm predictive power

The algorithm predictive power assesses the probability that a threat is actually in the category i for $i = H, M, L$ (“High”, “Medium” or “Low”) if the algorithm classifies it in the category i.

P_H is the probability that a threat is actually in category “High” if the model classifies it in category “High”, as given by Equation (11).

$$\mathrm{P}_H = \frac{T_H}{T_H + F_{H/M} + F_{H/L}} \tag{11}$$

P_M is the probability that a threat is in category "Medium" if the model classifies it in category "Medium" is given by Equation (12).

$$\mathrm{P}_M = \frac{T_M}{T_M + F_{M/H} + F_{M/L}} \tag{12}$$

P_L is the probability that a threat is in category "Low" if the model classifies it in category "Low", as given by Equation (13).

$$\mathrm{P}_L = \frac{T_L}{T_L + F_{L/H} + F_{L/M}} \tag{13}$$

These MoPs vary in [0, 1]. A value of 1 is indicative of high predictive power resulting in a better performance whereas a value of 0 is indicative of low predictive power (and then low performance).

Kappa Statistic

The Kappa statistic is a statistical measure that assesses the proportion of specific agreement taking into consideration the improvement over chance [14], [15]. It is generally thought to be a more robust measure than correct classification rate since kappa takes into account the agreement occurring by chance.

Kappa K is the proportion of specific agreement between the algorithm and the reality, as given by Equation (14).

$$\mathrm{K} = \frac{A - E}{T - E} \tag{14}$$

where A is the number of time the algorithm agreed with the reality, E is the number of time the classification by chance would have agreed with reality and T is the total number of cases.

[14] proposed a scale to assess the agreement with Kappa statistics. A Kappa value under 0.4 indicates poor agreement, whereas a value above 0.4 is indicative of good agreement.

Table 3: Scale for assessing the agreement with Kappa Statistics [14]

Kappa	Agreement
0.00 - 0.05	None
0.05 - 0.20	Very poor
0.20 - 0.40	Poor

0.40	-	0.55	Moderate
0.55	-	0.70	Good
0.70	-	0.85	Very good
0.85	-	0.99	Excellent
0.99	-	1.00	Perfect

4.2.2 Agreement measures for threat ranking algorithm

The threat ranking algorithm provides a ranking of threats inside each category. In order to evaluate the performance of this ranking, we propose to measure the agreement between the ranking given by the algorithm and the actual ranking given by the SME. The rank correlation coefficient introduced by [16] and then extended by [17] will be used. This rank correlation coefficient allows comparing two weak orderings with ties.

A weak order of n objects will be represented using the $n \times n$ score matrix $\{ a_{ij} \}$ as follows.

$a_{ij} = 1$ if threat i is ranked ahead of or tied with threat j.

$a_{ij} = -1$ if threat i is ranked behind threat j.

$a_{ij} = 0$ if $i = j$.

Let $\{ \rho_{ij} \}$ the weak order representing the ranking given by the algorithm and $\{ R_{ij} \}$ the weak order representing the actual ranking. The rank correlation between the ranking ρ given by the algorithm and the actual ranking R given by the SMEs is given by the dot product of their score matrices [17], as calculated in Equation (15).

$$\tau_x(\rho, R) = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} \rho_{ij} . R_{ij}}{n(n-1)} \tag{15}$$

The τ_x coefficient is assigned values in [-1,+1]. τ_x is assigned +1 when the ranking given by the algorithm is in full agreement with the actual ranking and -1 when the ranking given by the algorithm is in full disagreement with the actual ranking.

4.2.3 Timeliness

Timeliness measures consist of evaluating if the threat classification and ranking algorithms provide or not the response at time.

Let us note that threat classification and ranking algorithms compute a first result and then continuously re-evaluate it. Consequently, time at which the threat is first classified (or ranked)

corresponds to the execution time of the algorithm until a first classification is provided. Unless the first classification is not the actual classification, the algorithm will continue re-evaluating the results until reaching the correct classification. Once the correct classification is reached, no more change in the results will occur.

Let us first present some notations.

T is the latest time at which the results of the threat evaluation (including classification and ranking) is of value for the TE process.

T_C is the time (since detection) at which the threat is first classified. In other words, it corresponds to the time at which the algorithm produces its first classification results.

T_{CC} is the time (since detection) at which the threat is correctly classified. In other words, it corresponds to the time at which the algorithm produces the correct results (after eventually some iterations).

T_R is the time (since detection) at which the threat is first ranked. In other words, it corresponds to the time at which the algorithm produces its first ranking results.

T_{CR} is the time (since detection) at which the threat is correctly ranked. In other words, it corresponds to the time at which the algorithm produces the correct results (after eventually some iterations).

For the threat classification algorithm, we propose the following time-related metrics.

τ_C: Elapsed time before the threat is first classified over the latest time T at which the TE is of value as calculated in Equation (16).

$$\tau_C = \frac{\mathrm{T}_C}{\mathrm{T}} \tag{16}$$

τ_{CC}: Elapsed time before the threat is correctly classified over the latest time T at which the TE is of value as calculated in Equation (17).

$$\tau_{CC} = \frac{\mathrm{T}_{CC}}{\mathrm{T}} \tag{17}$$

τ_{EC}: Time to recover from error (time during which threat classification is incorrect over the latest time T at which the TE is of value as calculated in Equation (18).

$$\tau_{EC} = \frac{\mathrm{T}_{CC} - \mathrm{T}_{C}}{\mathrm{T}} \tag{18}$$

The same timeliness measures than for the classification algorithms could be used to assess the timeliness of the threat ranking algorithm.

τ_R: Elapsed time before the threat is first ranked over the latest time T at which the TE is of value as calculated in Equation (20).

$$\tau_{R} = \frac{\mathrm{T}_{R}}{\mathrm{T}} \tag{19}$$

τ_{CR}: Elapsed time before the threat is correctly ranked over the latest time T at which the TE is of value as calculated in Equation (20).

$$\tau_{CR} = \frac{\mathrm{T}_{CR}}{\mathrm{T}} \tag{20}$$

τ_{ER}: Time to recover from error (time during which threat ranking is incorrect over the latest time T at which the TE is of value as calculated in Equation (21).

$$\tau_{ER} = \frac{\mathrm{T}_{CR} - \mathrm{T}_{R}}{\mathrm{T}} \tag{21}$$

4.2.4 Sensitivity analysis

Sensitivity analysis consists of evaluating if small changes in the values of inputs have an impact on the results of the algorithms. Small changes in the values of inputs will correspond to a noise in the values of inputs variables. For threat classification and ranking algorithms, it is preferable that small changes in input (due to uncertainty or errors of measurement) have no impact on the result because of the discrete nature of these results. We propose to conduct sensitivity analysis in order to: 1) verify if small change in the values of inputs has or not an impact on both the category to which each threat is assigned and the ranking of the threat inside the category and 2) identify the bounds of the interval in which the inputs could vary without changing the final results of the algorithm. The objective of this analysis is to verify if noise in data will affect the final results.

4.2.5 Input degradation analysis

In this section, we are concerned with measuring the variation of the results of threat classification and ranking algorithms when the quality of their input decreases. We propose to apply the degradation concept introduced by [18]. The approach of "degradation" analyzes how system output changes as a function of degrading system input. In the work of [18], the degradation of the input quality is considered in term of incompleteness and incorrectness. The results in term of output quality are measured using accuracy measures. The degradation study concludes that the system is robust if it produces an output with monotonically decreasing quality as function of decreasing input quality. In fact, according to [18], "when input quality decreases, a system that produces an output with a fluctuating quality is much less predictable than a system that produces an output with a monotonically decreasing quality". In the following, we explain how the degradation approach will be used for the threat classification and ranking algorithms.

Degradation of the input quality

For threat classification and ranking algorithms, when information regarding an input is not available, a default value is assigned to that input. So in our analysis, the degradation of the input quality will correspond to the substitution of the input actual value with the default value. The degradation is more accentuated depending on the number of inputs for which default values are assigned. The more the volume of the inputs that are assigned default values and the more the input quality is degraded. Input quality varies in [0, 1]; 0 if default values are assigned to all inputs and 1 if there is no degradation in the inputs. The inputs that could be considered are: kinematics information (range, speed, bearing, and azimuth), Closed Point of Approach (CPA), Time to Closest Point of Approach (TCPA), lethality, identity, type, behaviour and opportunity.

Effects of the input degradation on the performance of TE algorithms

The degradation study consists of degrading the inputs and measuring the performance of the algorithms with the degraded set of inputs. The performance of the threat classification algorithm is measured using the accuracy measures presented in Section 4.2.1. The performance of the ranking algorithm is measured using the rank correlation coefficient formulated in Section 4.2.2. The performance of the algorithm is expected to decrease over a degradation of the input quality as shown in Figure 11. Table 4 presents the expected variation of each MoPs with the input quality degradation.

Table 4: Expected variation of MoPs over a degradation of the input quality

MoPs		**Expected variation with the input quality degradation**
Threat classification algorithm		
ρ	Misclassification rate	Increase
S_i	Sensitivity of class i	Decrease
$\varepsilon_{i/j}$	Algorithm errors (Type 1, Type 2)	Increase
P_i	Algorithm predictive power	Decrease
K	Kappa statistic	Decrease
Threat ranking algorithm		
τ_x	Rank correlation coefficient	Decrease

In other words, the more the volume of inputs - for which default values are assigned- is important and the more the accuracy of the threat classification algorithm and the agreement of the threat ranking algorithm are expected to decrease. Degradation analysis is a complementary technique in the assessment of the algorithm performance. It is concerned with the behaviour of the algorithm results. The idea behind the degradation study is that the actual values of inputs should give better results than the default values. The algorithm will have a good performance if the quality of its results follows a monotonically decreasing function (for different scenarios in different contexts) as shown in Figure 11 and Figure 12. Degradation analysis will consist of a series of graphs which represent MoPs over the volume of inputs for which default values are assigned. For each scenario, there will be as many graphs as performance measures presented in Section 4.2.1 and Section 4.2.2. Figure 11 and Figure 12 show the desired shape of the MoPs as a function of the input quality.

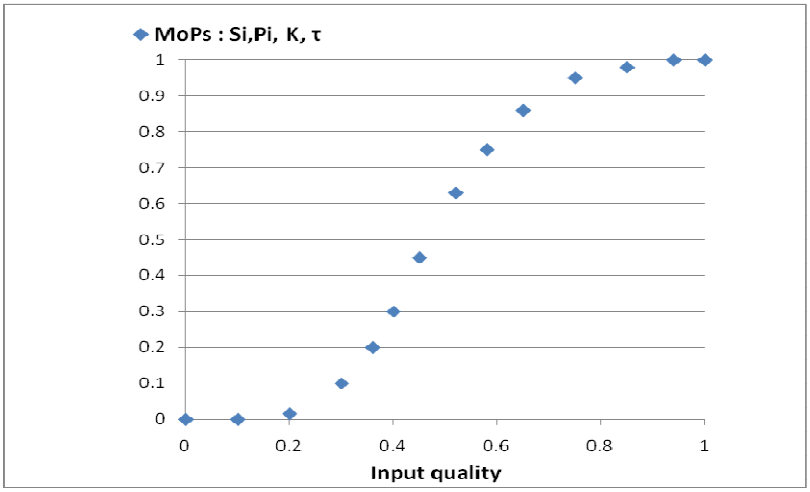

Figure 11: Algorithm's sensitivity, predictive power, Kappa statistic and rank correlation coefficient as a function of the input quality

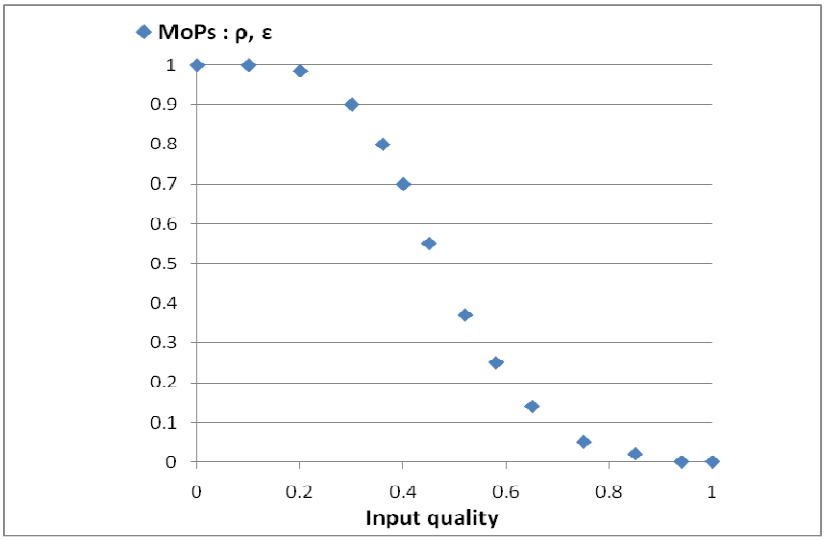

Figure 12: Algorithm's errors and misclassification rate as a function of the input quality

4.2.6 Anytime behaviour analysis

Because of the necessity to respond in time to instantaneous threats and changing situations, the TE algorithms should give recommendations under tight temporal constraints. Consequently, the behaviour of the algorithms toward the runtime issues should be considered in the performance evaluation. In this section, our focus will be on evaluating the anytime aspects of the threat classification and ranking algorithms.

Desired properties to be verified

Threat classification and ranking algorithms, as any other anytime algorithm, should have correct behaviours toward the runtime. Thus, it is important to show that the algorithm is a well-behaved anytime algorithm. To do so, we refer to a set of desired properties proposed in [19]. These properties are presented below.

- Measurable quality: the quality of the algorithm's results is measurable as long as the correct result can be determined
- Monotonicity: the quality of the result is a nondecreasing function of time and input quality.
- Consistency: the algorithm should reach the same outcome quality for a given amount of time and input quality. If algorithms do not guarantee a deterministic output quality for a given amount of time, it will be important to have a narrow variance.
- Diminishing returns: the improvement in results quality is greater at the early stages of the computation, and it decreases over time.
- Interruptibility: the algorithm can be stopped at anytime and provide some answer.
- Preemptability: the algorithm can be suspended and resumed with minimal overhead.

Performance profile analysis

In order to assess the anytime behaviour of threat classification and ranking algorithms, the algorithm's performance improvements over time are summarized in charts called performance profile (PP) [19]. Figure 13 and Figure 14 show the performance profile shape that a well-behaved anytime algorithm should have.

To provide these graphs, the performance of the threat classification algorithm is measured using the accuracy measures presented in Section 4.2.1 and the performance of the ranking algorithm is measured using the rank correlation coefficient formulated in Section 4.2.2. The accuracy of the threat classification algorithm and the agreement of the threat ranking algorithm are expected to increase with more available runtime. The algorithm will have a good performance profile if the quality of its results (for different scenarios in different contexts) is a monotonically increasing function with diminishing return as in Figure 13 and Figure 14. Diminishing returns refer to the fact that the improvement in output quality is greater at the early stages of the computation, and decreases over time.

Performance profile analysis will consist of graphs which represent MoPs over the volume of inputs for which default values are assigned. For each scenario, there will be as many graphs as MoPs presented in Section 4.2.1 and Section 4.2.2. Figure 13 and Figure 14 show the desired shape of the MoPs as a function of the execution time.

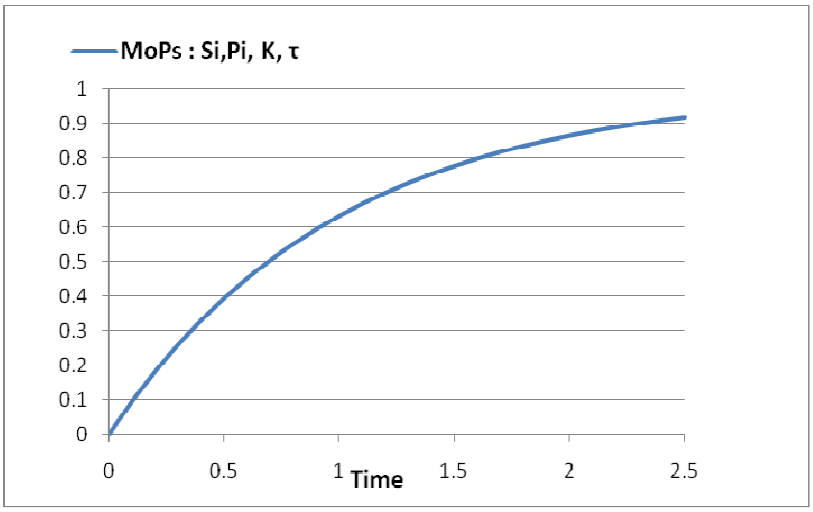

Figure 13: Performance profile (algorithm's sensitivity, predictive power, Kappa statistic and rank correlation coefficient as a function of execution time)

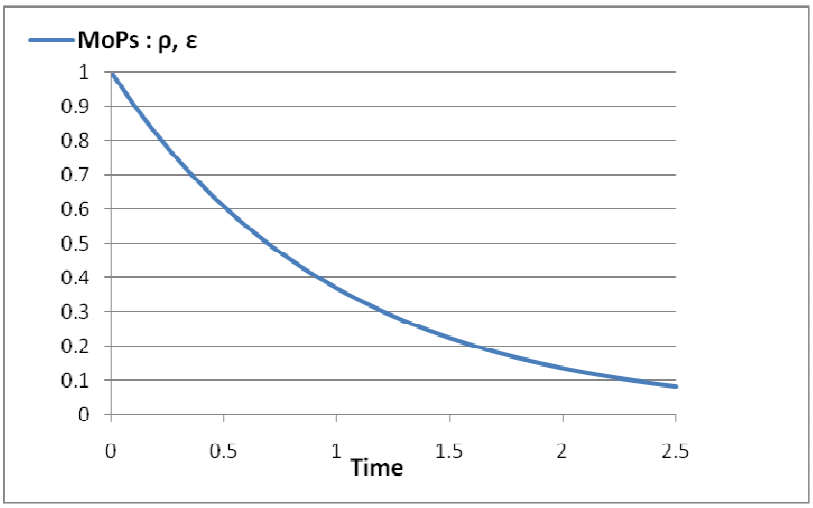

Figure 14: Performance profile (algorithm's errors and misclassification rate as a function of execution time)

4.3 Human factors metrics

In order to assess the human-factors related performance, we propose to evaluate the level of cognitive activity that the human operators will be experiencing in the context of the human-in-the-loop experiments. Evaluating the level of cognitive activity that the human operator is experiencing is crucial. This evaluation verifies if higher quality of decision making is accompanied or not by an overwhelming of the operator.

The Cognitive Systems Engineering (CSE) framework ([20], [21]) will be used as a basis for this evaluation (Figure 15). The CSE framework presents how we can measure the overall level of cognitive activity, referred to as the Dynamic Mental Model. The Dynamic Mental Model

(theoretical construct) is explained by three empirical constructs: situation awareness, workload, and task-relevant behaviour. These empirical constructs will be considered to measure the operator performance based on the assumption that higher situation awareness, lower workload and higher task-relevant behaviour will imply higher operator performance.

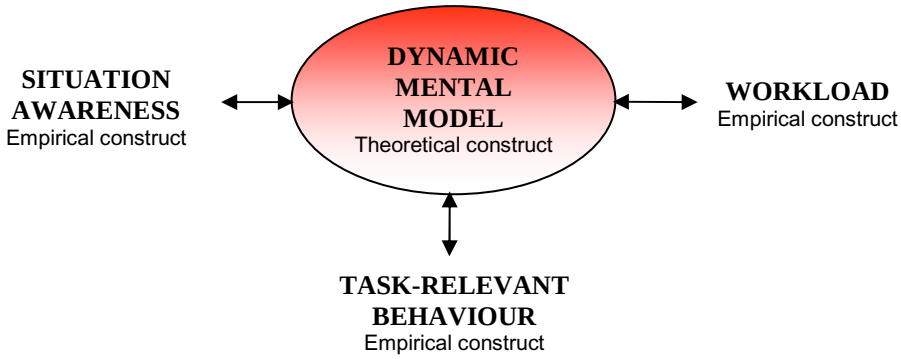

Figure 15: Cognitive systems engineering framework [20]

In this section, the operator's measures of performance are briefly presented. A more detailed presentation of these measures is provided in the demonstration and experimentation plan [22].

Situation Awareness (SA): SA can be briefly summarized as, "Knowing what is going on around you" [23]. Three levels of SA have been identified: (1) perception and selective attention, (2) comprehension, and (3) prediction [23]. Perception and selective attention refers to the ability to distinguish between important and non-important information in the environment. Comprehension is the ability to integrate and comprehend sources of information. Prediction is the ability to predict the future status of elements and events in the environment based on the awareness of their current state. A high level of SA is necessary to reliably make accurate decisions. Different ways of measuring SA are presented in Annex C.

Workload: Workload measures of cognitive activity are based on the premise that humans are limited in their ability to process information and to respond accurately and appropriately. In other words, there is a limited pool of cognitive resources available to operators at any given moment in time. Once this pool is depleted (due to high workload), responses to each additional task are compromised [22]. Different measures of workload are provided in Annex C.

Task-relevant behaviour: Task-relevant behaviour measures refer to the behaviour and actions of the operator that are directly affected by the TE capability. [22]. Measures of task relevant behaviour are provided in Annex C.

The TE capability, if effective, will result in increasing an operator's SA, decreasing operator workload and will ensure that the right information is presented at the right time (and in the right format). All this factors will result in increasing the CDSC-operator decision making quality.

Let us note that the operators that will participate in the experiments should have similar experience. This is very important to avoid the case where the level of cognitive activity is higher because of the experience of the operator.

4.4 Measures of TE decision-making quality

Accuracy measures: The accuracy of TE decisions made by the operator when using the TE capability should be evaluated in order to see if they are sufficiently correct. Accuracy is computed by comparing the decision made by the operator when using the TE capability to the desired decision. SMEs will intervene in this evaluation process a priori by defining what constitutes a sufficiently correct decision given the context and the situation.

Measures of accuracy that will be used to evaluate quality of decisions made by the operator are very close in their formulation to the measures of accuracy and agreement presented in Section 4.2.1 and Section 4.2.2. As for MoPs for the algorithms, confusion matrix is used. The assessed category will correspond here to the classification and ranking decisions made by the operator when using the TE capability. The actual category is the category proposed by the SMEs.

Agreement measures: As in Section 4.2.2, we propose the use of the rank correlation coefficient presented in Equation (15) to measure the agreement between the ranking given by the algorithm and the actual ranking given by the SME.

Timeliness: In dynamic situations, time is of high value and can dictate not only the eventual outcome of the decision but also the operator's pattern of decision making. Operators should be able to quickly react. The speed of decision making is couched in terms of decision timeliness. Measures of timeliness presented in Section 4.2.3 are still applicable in this case. The only difference is that the classification and ranking here are made by the system (operator using the TE capability) instead of by the algorithms as it was the case earlier in Section 4.2.3.

Consistency: It refers to the fact that operators who have made accurate situation assessment and correct inferences based on given data, should reach the same outcome each time a similar decision is made based on the same data. The consistency of decision making is a useful insight into decision making quality. In fact, differences of outcome from decisions based on the same data could indicate inappropriate or incorrect reasoning processes. To measure consistency, similar scenarios will be proposed to the operator on two separate occasions at different moments in time. This measure aims at demonstrating the extent to which the operator is able to produce stable and consistent decisions across time.

Finally, let us note that as for measuring the algorithms performance, measuring decision-making quality will be done in two steps. In the first step, we consider a scenario with a large number of threats, which are already classified and ranked by one (or a group of) SME(s). These threats are classified and ranked by the operator when using the TE capability and results are compared with the SME classification and ranking. In the second step, we consider other scenarios (in different context) that have different level of complexity. For each scenario, we compare again the classification and ranking of threats by the operator using the TE capability with the classification and ranking of the SMEs. This context-based analysis aims at verifying that the quality of the decisions remains high whatever the scenario and whatever the context.

5 Conclusion

In this document, we were concerned with the TE capability developed within the INCOMMANDS TDP to improve TE decision-making. The general objective of this work was to provide measures of performance for the evaluation of this capability as well as the evaluation of the TE decisions made by the operator when using this capability. A conceptual model explaining how these evaluations will be done was proposed. First, we formulated a set of MoPs for the assessment of the quality of the recommendations provided by the TE capability. These measures are: measures of accuracy and agreement, measures of timeliness, sensitivity analysis, input degradation analysis and anytime behaviour analysis. Second, we considered that the assessment of the level of cognitive activity that the human operator will be experiencing is important. This evaluation will assess if higher quality of decision making is reached while the operator is overwhelmed or not. To do so, we proposed the use of the following MoPs for the operator: situation awareness, workload and task relevant behaviour. Third, the quality of the TE decisions made by the operator (when using the capability) will be measured using complementary MoPs such as accuracy, timeliness and consistency. When analyzed together, these measures provide converging evidence that decision making quality has either increased or decreased.

Concerning the INCOMMANDS evaluation and experimentation, only human factor evaluation that aims at evaluating the human factors aspects of the INCOMMANDS system was carried out. It consisted of a heuristic Evaluation by a Human Factors analyst and usability and utility testing with naval operators [4]. The evaluation of the algorithms and the assessment of the decision quality did not take place during the course of the project because of temporal and budget constraints. However, even if the measures of performance proposed in this work were thought originally in the spirit of evaluating the TE capability within the INCOMMANDS TDP, the results presented in this work are widely applicable for other similar problems. For instance, they could be used in future maritime projects as the Coalition Maritime Missile Defence (CMMD) TDP. Also, they could be used to evaluate other capabilities based on classification of items and for which the human operator intervene in the final decision (ex: object recognition and identification process). In fact, the conceptual model could be used in any other context where a capability supports a human operator in his decision-making process. The idea behind is to evaluate the quality of the decisions made by the operator simultaneously with evaluating his performance and the quality of the capability recommendations. Second, measures of accuracy proposed to evaluate the quality of the recommendations provided by the capability remain appropriate for other capabilities that consist of classification of items inside categories (in particular the three-class classification problems). Also, measures of agreement for ranking algorithms remain suitable to compare two rankings of items in any other context. As well, sensitivity analysis, input degradation analysis and anytime behaviour analysis are sufficiently general to be used in other situations.

This page intentionally left blank.

References

[1] Paradis, S., Benaskeur, A., Oxenham, M. and Cutler, P., (2005) "Threat Evaluation and Weapons Allocation in Network-Centric Warfare" In Proceeding of 8th International Conference of Information Fusion.

[2] Hampton, D.R. (1998), "Command and Control in Joint Operations: Separate Functions, their Purpose, and Application to Battle Command in the 21st Century", Thesis, Faculty of the Naval War College.

[3] Benaskeur, A. and Kabanza, F. (2008) "Combat Power Management for INCOMMANDS TDP: Characterization of the Problem and Review of Applicable Technologies", DRDC Valcartier TR 2008-286.

[4] Baker, K., Hagen L., (2009), INCOMMANDS TDP: Human Factors Evaluation of the Command Decision Support Capability Prototype, DRDC Toronto CR 2009-041.

[5] Waltz, E., and Llinas, J., (1990), Multisensor Data Fusion, Artech House, Norwood, MA.

[6] Paradis, S., Pageau, N., Belhumeur C., and Duclos-Hindié N., (2002), "Analysis of a Threat Evaluation Algorithm Modification Proposal for Halifax Class CCS 330", DRDC Valcartier ECR-238, September 2002, 22 pages, CONFIDENTIAL.

[7] Roy, J., Paradis, S., and Allouche, M. (2002), "Threat Evaluation for Impact Assessment in Situation Analysis Systems", SPIE Proceedings, Vol. 4729, Signal Processing, Sensor Fusion, and Target Recognition XI, Orlando, 1-5 April 2002, 14 pages.

[8] Herbst, A., (2009) "INCOMMANDS Sea Trials Command Decision Support Lab Prototype: System Requirement Specification", DRDC Valcartier CR 2009-058.

[9] Oxenham, M. (2003) "Using Contextual Information for Extracting Air Target Behaviour from Sensor Tracks", In Proceedings of the Signal Processing, Sensor Fusion, and Target Recognition, XII Conference at the 17th SPIE Annual AeroSense Symposium.

[10] Prevost F., Fawcett, T., Kohavi, R. (1998), "The Case Against Accuracy Estimation for Comparing Induction Algorithms", Proceeding of the 15th International Conference on Machine Learning, Morgan Kaufmann, San Francisco, CA, pp. 445-453.

[11] Bradley, A. (1997) "The Use of the Area under the ROC Curve in the Evaluation of Machine Learning Algorithms" Pattern Recognition, vol 30, pp. 1145-1159.

[12] Adams, N.M., and Hand, D.J., (2000), "Improving the practice of classifier performance assessment", Neural Computation, vol 12, pp 305-311.

[13] Metz, C. (1978), "Basic principles of ROC analysis", Seminars in Nuclear Medicine, vol 8, no 4, pp 283-298.

[14] Monserud, R.A., and Leemans, R., (1992), “Comparing Global Vegetation Maps with the Kappa Statistic”, Ecological Modelling, vol 62, pp 275-293.

[15] Landis, J. R. and Koch, G.C. (1977), “The measurement of observer agreement for categorical data”, Biometrics 33, pp. 159-174.

[16] Kendall M. (1948), Rank Correlation Methods, Charles Griffin and Company Limited: London.

[17] Emond, E.J. & Mason, D.W, (2002) “A New Rank Correlation Coefficient with Application to the Consensus Ranking Problem”, Journal of Multi-Criteria Decision Analysis, 11, 17-28.

[18] Groot, P., ten Teije A., van Harmelen F., (2003), “A Quantitative Analysis of the Robustness of Knowledge-Based Systems through Degradation Studies”. Knowledge and Information Systems, vol 7, no 2, pp. 224-245.

[19] Zilberstein, S., (1996), “Using anytime algorithms in intelligent systems”, American Association for Artificial Intelligence, 0738-4602, Fall 1996.

[20] Paradis, S., Elm, W.C., Potter, S.S. Breton, R. and Bossé, É., (2002), “A Pragmatic Cognitive System Engineering Approach to Model Dynamic Human Decision-Making Activities in Intelligent and Automated Systems”, Proceedings of the NATO RTA HFM Symposium on The Role of Humans in Intelligent and Automated Systems, Warsaw, Poland.

[21] Elm, W.C. Potter, S.S. Gualtieri, J.W., Roth, E.M. and Easter, J.R., (2003), “Applied Cognitive Work Analysis: A Pragmatic Methodology for Designing Revolutionary Cognitive Affordances”, in Hollnagel, E. (Ed.), Handbook of Cognitive Task Design, Mahwah, NJ: Erlbaum.

[22] Scott Brown M., Herdman C.M. (CogSim Ltd) (2006) “INCOMMANDS Spiral 2: Demonstration and Experimentation Plan”, DRDC Valcartier CR 2006-577.

[23] Endsley, RM (2000), “Direct Measurement of situation awareness: validity and use of SAGAT, in Situation Awareness and Management”, Endsley MR. and DJ. Eds, Lawrence Erlbaum Associates, Mahwah, N.J., pp 147-173.

[24] Zimmermann, N.E., Simtest, http://www.wsl.ch/staff/niklaus.zimmermann/programs/progs/simtest.pdf

[25] Fielding, A.H., and Bell J.F. (1997), “A review of methods for the assessment of prediction errors in conservation presence/absence models”, Environmental Conservation, vol 24, no 1, pp 38-49.

Annex A Default criteria used in the TE capability

A.1 Default criteria for reactive and VOI tests

Table 5: Reactive criteria [8]

Criteria ID	Criteria	Explanation
A	0 meters < Track Altitude < 100 meters	Sea skimming profile
B	Track Speed > 250 meters/second	Suggestive of Anti-Ship Missile (ASM)
C	Track CPA to Asset Point < 2000 meters	Incoming trajectory
Specification of Reactive Test: A track is deemed to be a reactive track if: [(A = TRUE) AND (B = TRUE) AND (C = TRUE)]		

Table 6: Volume of interest criteria [8]

Criteria ID	Criteria	Explanation
A	Track Range to Asset Point < 250,000 meters	Close to the asset point
Specification of VOI Test: A track is deemed to be within the VOI if: [A = TRUE]		

A.2 Default criteria for "determining hostile intent" sub-function

Table 7: Hostile Intent Criteria [8]

Criteria ID	Criteria	Explanation
A	Any ASM that appears in the environment of the asset point can be determined to exhibit hostile intent in respect to the asset point.	There is no alternative explanation for the appearance of an ASM.
B	If an aircraft has been determined to have launched an ASM, this is an item of evidence supporting the establishment of hostile intent on the part of the aircraft.	
Specification of overall criteria for determining the presence of hostile intent: • Satisfaction of A suffices to establish hostile intent of an ASM. • Satisfaction of B suffices to establish hostile intent of an aircraft.		

A.3 Default criteria for "determining capability" sub-function

Within the CDSC, the only capability determinations supported are those in respect to threat scenarios consisting of:

1. A fighter aircraft equipped with target acquisition radar and carrying one or more ASM of the same type, each equipped with the same kind of active radar seeker.

2. An ASM in flight, equipped with an active radar seeker.

This section presents a summary of the default criteria for establishing the presence of opportunity in respect to the two threat types identified above.

Table 8: Establishing opportunity for an aircraft with an ASM [8]

Criteria ID	Criteria	Explanation
A	The aircraft has > 0 ASMs onboard.	Aircraft has to be carrying at least one missile.
B	The distance between the aircraft and the asset point is less than or equal to the maximum detection range of the aircraft's targeting radar.	The aircraft must have the ability to sense (detect) the asset point.
C	The asset point is within the effective range of the ASM.	The ASM that might be launched must have sufficient energy to reach the asset point if it is launched.
D	Based on the assumption that the ASM, once launched, will travel in a straight line following the same heading as the aircraft that carries it, the asset point, at some time in the future, lies within the seeker cone (window in angle as limited by maximum range).	The asset point must, at some point in time, lie within the seeker cone of the ASM so that the asset point has even the possibility of being sensed (detected).
A track representing an aircraft with ASMs is deemed to have opportunity to harm the asset point if: [(A=True) AND (B=True) AND (C=True) AND (D=True)] If the conditions to establish the presence of opportunity are not satisfied, the threat will be deemed to not have the opportunity to harm the asset point.		

Table 9: Establishing opportunity for an ASM in flight [8]

Criteria ID	Criteria	Explanation
A	The asset point is within the effective range of the missile; conditioned on the assumption that missile launch is temporally coincident with the creation of the associated system track.	The ASM must have sufficient energy to reach the asset point if it is launched.
B	Based on the assumption that the ASM will travel in a straight line following a non-varying heading, the asset point, at some time in the future, lies within the seeker cone (window in angle as limited by maximum range).	The asset point must, at some point in time, lie within the seeker cone of the ASM so that the asset point has even the possibility of being sensed (detected).
C	The time to go until the ASM attains CPA with respect to the asset point is below a specified value and the asset point lies within the seeker cone at this time and the threat seeker is not on.	Determination that an ASM malfunction has occurred. This criterion deals with operational capability. It is assumed that the ASM's seeker needs to illuminate its target for some minimum time prior to attaining CPA in order to have any chance of damaging the target.
A track representing an ASM in flight is deemed to have opportunity to harm the asset point if: [(A=True) AND (B=True) AND (C=False)] If the conditions to establish the presence of opportunity are not satisfied, the threat will be deemed to not have the opportunity to harm the asset point.		

This page intentionally left blank.

Annex B Performance metrics for binary classification algorithms

Performance metrics to assess the quality of binary classification algorithm's results are presented in this annex. In particular, measures presented here are based on confusion matrix. Table 10 presents the confusion matrix for binary classification. T_p and T_n refer respectively to the true positive cases (assessed positive by the algorithm and they are actually positive cases) and true negative cases (assessed negative by the algorithm and they are actually negative cases). F_p and F_n refer respectively to false positive cases (assessed positive by the algorithm and they are actually negative cases) and false negative cases (assessed negative by the algorithm and they are actually positive cases).

Table 10: Confusion matrix for binary classification [11]

	Actual Positive	**Actual Negative**
Assessed Positive	T_p	F_p
Assessed Negative	F_n	T_n

Correct classification rate and misclassification rate are the simplest measures of accuracy. They are not very relevant if the prevalence of the data set is very low. Sensitivity (resp. specificity) measures the proportion of positive (resp. negative) cases that are correctly classified. Positive predictive power (resp. negative predictive power) measures the probability that the case is observed as positive (resp. negative) if the model classifies it as positive (negative). False positive rate (resp. false negative rate) measures the proportion of positive (resp. negative) cases that are incorrectly classified. According to [24], the error measured by the false positive rate is called "commission or type I error", whereas the error measured by the false negative rate is called "omission or type II error".

Table 11: Performance metrics for binary classification algorithms [24],[25]

Metric	**Definition**	**Equation**
Corect classification rate	Proportion of cases that are correctly classified	$\frac{T_p + T_n}{T_p + F_n + T_n + F_p}$
Misclassification rate	Proportion of cases that are incorrectly classified	$\frac{F_p + F_n}{T_p + F_n + T_n + F_p}$
Sensitivity	Proportion of positive cases that are correctly classified	$\frac{T_p}{T_p + F_n}$

Specificity	Proportion of negative cases that are correctly classified	$\frac{T_n}{T_n + F_p}$
Positive predictive power	Probability that a case is observed as positive if the model classifies it as positive	$\frac{T_p}{T_p + F_p}$
Negative predictive power	Probability that a case is observed as negative if the model classifies it as negative	$\frac{T_n}{T_n + F_n}$
False positive rate	Proportion of positive cases that are incorrectly classified	$\frac{F_p}{T_n + F_p}$
False negative rate	Proportion of negative cases that are incorrectly classified	$\frac{F_n}{T_p + F_n}$
Prevalence	Proportion of actual positive cases.	$\frac{T_p + F_n}{T_p + F_n + T_n + F_p}$

Annex C Measures of operator performance

The content of this Annex is extracted from the Demonstration and Experimentation Plan (DEP) document [22] prepared for the INCOMMANDS TDP.

C.1 How to measure situation awareness?

Situation Awareness Global Assessment Technique (SAGAT): Traditionally, the SAGAT is administered by freezing the simulation and blanking the screens at various predetermined parts of the scenario. At this point the operators would be asked to reproduce their mental representation for the location and type of entities/contacts/threats. Unfortunately, the intrusiveness of this approach often results in the crew becoming less immersed in the scenario and therefore leads to it being perceived as less realistic. Alternatively, the SAGAT can be administered after a scenario has been completed, so long as the lag between the end of the scenario and the SAGAT is kept to a minimum. It is anticipated that each operator will be given a multiple-choice SAGAT test where five PPI displays are shown, but only one of them will be the correct representation of what the PPI display looked like for a given part of the scenario. The foils (i.e., incorrect options) will be created by (individually or in combination) adding/deleting contacts from the PPI display or by changing the type of contact (e.g., from hostile to friendly).

Situation Present Assessment Method (SPAM): This method is similar to the SAGAT in that it provides an objective measure of SA. However, it differs from the SAGAT by dissociating the effects of memory from the measurement of SA per se. Unlike the SAGAT, which requires operators to explicitly remember the location/status of objects, the SPAM makes allowance for the fact that even though an operator may not have an explicit memory for a piece of information, they may know exactly where to look for it. In this sense, an operator is deemed to have good SA of something, so long as they know where to find it, and not necessarily because they have it stored in memory.

Change Blindness: This measures the operator's ability to notice changes (i.e., freezing and changed entities) on their displays. It is anticipated that crews will often fail to notice these changes; however, in cases where the changes are noticed, the time required to notice them will be measured. Change blindness is designed to compliment the SAGAT and the SPAM measures by providing an on-line metric of the crew's SA for abrupt changes to or freezing of their physical displays. Unlike the SAGAT, which, in this instance, measures individual SA, this measure will index the crew's collective SA. At various pre-determined points in the scenario, the PPI will either freeze or instantaneously change (e.g., add/remove a contact, change the status of a contact). This freezing/change will persist for a pre-determined period of time, after which the PPI display will return to normal. In cases where the crews notice this freezing/change, their response times (i.e., the difference in time between the onset and awareness of the freezing/change) will be recorded. Cases where crews will fail to notice this freezing/change will be logged as a "miss."

Subjective Assessment of SA: Each operator will be given a questionnaire that asks them to rate their SA on multiple tasks on a seven-point Likert scale where 1 represents "very low" SA and 7 represents "very high" SA. Operators will circle a number from one to seven that they feel best

represents their SA for that task. The SA questionnaires will be tailored such that operators are only asked to provide ratings for relevant tasks.

C.2 How to measure workload?

Task Load Index (TLX): This measure is derived from the NASA TLX Workload assessment scale and will provide a method for evaluating operator workload. In addition to the six workload subscales used in the standard NASA TLX, the version of the TLX used here will also include a seventh subscale that will measure the operator's overall workload. Specifically, operators will be asked to rate their workload for:

1. Mental Demand
2. Physical Demand
3. Temporal Demand
4. Frustration Level
5. Effort
6. Performance
7. Overall

Operators will indicate their workload by placing an "X" on a horizontal scale where the leftmost marker represents 0% (i.e., very low) workload and the rightmost marker represents 100% (i.e., very high) workload. Each 10% interval is demarcated on the scale. If an "X" is placed in between two 10% demarcations (e.g., between 50% and 60%), then workload will be measured to the nearest 5% (i.e., 55%). Each of the seven workload subscales listed above will be rated separately. That is, operators will place an "X" on the Mental Demand scale (where 0% represents no mental demand and 100% represents very high mental demand). Then, they will place an "X' on the Physical Demand scale. This process will continue until they have placed an "X" on all seven subscales.

C.3 How to measure task-relevant performance?

To measures task-relevant performance, the following two measures are used.

Critical Task Sequence (CTS) completion time: CTSs will be defined a priori by SMEs and the experimental team and will represent a critical mission task that the crew must complete quickly and accurately in order to ensure ship survivability. Further these CTSs must have objective and easily observable start and end points that can be precisely captured by either the audio/video equipment or the scenario generation tools (e.g., button press responses, verbal commands). The time difference between the start and end points of each CTS will be calculated so as to determine the length of time required to complete the tasks.

Percentage of CTS shedding: The number of incomplete CTSs relative to the total number of CTSs will also be measured to compliment the CTS completion time data. It is standard practice to measure error rates when completion time (reaction time) for a task is a dependent variable. The rationale for doing so is to capture the fact that some participants will modify their response strategy under high workload conditions where they trade-off accuracy for speed. That is, they selectively respond to only a few tasks very effectively (i.e., low completion times and error rates), but at the cost of making errors on (or ignoring) other tasks. In the context of this experiment, the percentage of CTS shedding will effectively capture the strategy described above. That is, some operators may adopt a strategy where they only compete a select few of the CTSs (albeit effectively), but do so at the cost of shedding many other CTSs. Operators that use this strategy will score well in the CTS completion time measure, but will consequently score poorly for CTS shedding.

This page intentionally left blank.

List of symbols/abbreviations/acronyms/initialisms

ASMs	Anti-Ship Missiles
AWW	Above Water Warfare
C2	Command and Control
CDSC	Command Decision Support Capability
CDS Lab	Command Decision Support Laboratory
CMMD	Coalition Maritime Missile Defence
CPA	Closest Point of Approach
CPM	Combat Power Management
CTS	Critical Task Sequence
DND	Department of National Defence
DRDC	Defence Research & Development Canada
INCOMMANDS	Innovative Naval COMbat MANagement Decision Support
MoPs	Measures of Performance
OMI	Operator Machine Interface
PPI	Plan Position Indicator
R&D	Research & Development
ROC	Receiver Operating Characteristic
SA	Situation Awareness
SMEs	Subject Matter Experts
TCPA	Time to Closest Point of Approach
TDP	Technology Demonstration Project
TE	Threat Evaluation
TE/CPM	Threat Evaluation and Combat Power Management
TLX	Task Load Index
VOI	Volume Of Interest

This page intentionally left blank.

Glossary

Capability: refers to the ability of the threat to inflict injury or damage to the friendly forces, the assets they are protecting or the areas they are defending.

Intent: refers to the will or determination of the threat to inflict injury or damage.

Operator: A military person that has access to the functional and physical displays, receives recommendations from the TE capability and makes final TE decisions (classification and ranking of threats).

TE capability: A computer-based capability performing TE algorithms and presenting TE recommendations to the operator via functional OMI displays.

TE decisions: refer to the decisions made by the operator when using the TE capability.

TE recommendations: refer to the output of the TE algorithms implemented in the capability.

TE system: refers to the combination of the TE capability and the operator.

Incompleteness: a part of the data input or part of the knowledge used is missing. A data input could be a number of observations that might be hard or expensive to obtain.

Incorrectness: Data or knowledge is incorrectly represented. For example, incorrect data could be caused by a user who has made a wrong observation. Incorrect knowledge could be caused by faulty knowledge of an expert or a misunderstanding when coding the knowledge into the system.

This page intentionally left blank.

DOCUMENT CONTROL DATA

(Security classification of title, body of abstract and indexing annotation must be entered when the overall document is classified)

1. ORIGINATOR (The name and address of the organization preparing the document. Organizations for whom the document was prepared, e.g. Centre sponsoring a contractor's report, or tasking agency, are entered in section 8.)

Defence R&D Canada – Valcartier
2459 Pie-XI Blvd North
Quebec (Quebec)
G3J 1X5 Canada

2. SECURITY CLASSIFICATION (Overall security classification of the document including special warning terms if applicable.)

UNCLASSIFIED

3. TITLE (The complete document title as indicated on the title page. Its classification should be indicated by the appropriate abbreviation (S, C or U) in parentheses after the title.)

INCOMMANDS TDP: Measures of performance for the threat evaluation capability:

4. AUTHORS (last name, followed by initials – ranks, titles, etc. not to be used)

Frini, A.; Benaskeur, A.

5. DATE OF PUBLICATION (Month and year of publication of document.)

December 2009

6a. NO. OF PAGES (Total containing information, including Annexes, Appendices, etc.)

68

6b. NO. OF REFS (Total cited in document.)

25

7. DESCRIPTIVE NOTES (The category of the document, e.g. technical report, technical note or memorandum. If appropriate, enter the type of report, e.g. interim, progress, summary, annual or final. Give the inclusive dates when a specific reporting period is covered.)

Technical Memorandum

8. SPONSORING ACTIVITY (The name of the department project office or laboratory sponsoring the research and development – include address.)

Defence R&D Canada – Valcartier
2459 Pie-XI Blvd North
Quebec (Quebec)
G3J 1X5 Canada

9a. PROJECT OR GRANT NO. (If appropriate, the applicable research and development project or grant number under which the document was written. Please specify whether project or grant.)

9b. CONTRACT NO. (If appropriate, the applicable number under which the document was written.)

10a. ORIGINATOR'S DOCUMENT NUMBER (The official document number by which the document is identified by the originating activity. This number must be unique to this document.)

DRDC Valcartier TM 2009-240

10b. OTHER DOCUMENT NO(s). (Any other numbers which may be assigned this document either by the originator or by the sponsor.)

11. DOCUMENT AVAILABILITY (Any limitations on further dissemination of the document, other than those imposed by security classification.)

Unlimited

12. DOCUMENT ANNOUNCEMENT (Any limitation to the bibliographic announcement of this document. This will normally correspond to the Document Availability (11). However, where further distribution (beyond the audience specified in (11) is possible, a wider announcement audience may be selected.))

Unlimited

13. ABSTRACT (A brief and factual summary of the document. It may also appear elsewhere in the body of the document itself. It is highly desirable that the abstract of classified documents be unclassified. Each paragraph of the abstract shall begin with an indication of the security classification of the information in the paragraph (unless the document itself is unclassified) represented as (S), (C), (R), or (U). It is not necessary to include here abstracts in both official languages unless the text is bilingual.)

In this document, we are concerned with the Threat Evaluation (TE) capability, developed for the INCOMMANDS TDP to support the operator involved in the shipboard TE process. In particular, we focus on evaluating the recommendations provided by this capability as well as evaluating the final TE decisions made by the operator when using this capability. A conceptual model explaining how these evaluations will be done is proposed. First, we propose a set of measures of performance (MoPs) for the assessment of the quality of the recommendations provided by the TE capability. These measures are: measures of accuracy and agreement, measures of timeliness, sensitivity analysis, input degradation analysis and anytime behaviour analysis. Second, we consider that the assessment of the level of cognitive activity that the human operator is experiencing is important. This evaluation will assess if higher quality of decision making is reached while the operator is overwhelmed or not. To do so, we propose the use of the following human factors metrics: situation awareness, workload and task relevant behaviour. Third, the quality of the TE decisions made by the operator (when using the capability) will be measured using complementary MoPs such as accuracy, timeliness and consistency. When analyzed together, these measures provide converging evidence that decision making quality has either increased or decreased.

Dans ce document, nous nous intéressons à la nouvelle capacité d'évaluation des menaces développée dans le cadre du projet de démonstration technologique INCOMMANDS pour aider l'opérateur dans le processus d'évaluation des menaces. En particulier, nous nous concentrons sur l'évaluation des recommandations fournies par cette capacité ainsi que l'évaluation des décisions prises par l'opérateur en utilisant cette capacité. Un modèle conceptuel expliquant comment ces évaluations sont faites est proposé. Premièrement, nous proposons un ensemble de mesures de performance pour l'évaluation de la qualité des recommandations fournies par la capacité. Ces mesures sont : les mesures d'exactitude et d'accord, les mesures relatives aux temps d'exécution l'analyse de sensibilité, l'analyse de dégradation de la qualité des intrants et l'analyse du comportement "anytime" de l'algorithme. Deuxièmement, nous considérons que l'évaluation du niveau d'activité cognitive de l'opérateur est importante. Elle permettra d'évaluer si une meilleure qualité de la prise de décision est atteinte quand l'opérateur présente un niveau d'activité cognitif très élevé ou non. Ainsi, nous proposons l'utilisation des mesures de performance suivantes: éveil situationnel, charge de travail et comportements appropriés avec la tâche. Troisièmement, la qualité des décisions d'évaluation des menaces prises par l'opérateur sera mesurée en utilisant des mesures de performance complémentaires les unes aux autres telles que les mesures d'exactitude, les mesures relatives aux temps d'exécution et la consistance. Lorsque analysées ensemble, ces mesures multiples fournissent des preuves convergentes quant à l'amélioration ou non de la qualité de la prise de décision.

14. KEYWORDS, DESCRIPTORS or IDENTIFIERS (Technically meaningful terms or short phrases that characterize a document and could be helpful in cataloguing the document. They should be selected so that no security classification is required. Identifiers, such as equipment model designation, trade name, military project code name, geographic location may also be included. If possible keywords should be selected from a published thesaurus, e.g. Thesaurus of Engineering and Scientific Terms (TEST) and that thesaurus identified. If it is not possible to select indexing terms which are Unclassified, the classification of each should be indicated as with the title.)

Threat evaluation capability, accuracy measures, agreement measures, input degradation analysis, human factors.

Defence R&D Canada

Canada's Leader in Defence
and National Security
Science and Technology

R & D pour la défense Canada

Chef de file au Canada en matière
de science et de technologie pour
la défense et la sécurité nationale

www.drdc-rddc.gc.ca