COMMUNICATING UNCERTAINTY, ASSESSING INFORMATION QUALITY AND RISK, AND USING STRUCTURED TECHNIQUES IN INTELLIGENCE ANALYSIS

1.0 WORKSHOP OVERVIEW

In June 2016, at the Third Meeting of SAS-114, which was held at NATO STO's Centre for Maritime Research and Experimentation, Arne Biering offered to host the next meeting at the Danish Defence Intelligence Service in Copenhagen. As 2016 progressed, SAS-114 experienced further expansion. Kellyn Rein from Fraunhofer FKIE was appointed as a SAS-114 member from Germany, and Jonas Clausen Mork from the Swedish Defence Research Agency (FOI) was appointed as a member from Sweden, bringing the nation total to eight.

As well, over the latter half of 2016, defence intelligence stakeholder involvement expanded. Great Britain was planning to send Christina Clarke from DI Futures and Analytical Methods to update SAS-114 on recent developments in the analytic methods unit, and Alex Claver and Huib van de Meeberg were scheduled to speak about Netherlands' approach on various topics pertinent to SAS-114. Finally, some academic scholars had agreed to speak at the meeting. For instance, Floris Bex from the Department of Information and Computing Sciences at Utrecht University and Jonathan Nelson from the Center for Adaptive Behavior and Cognition at the Max Planck Institute for Human Development both agreed to join the meeting. When a tentative agenda was drafted, it became evident that the meeting had morphed into a full-fledged workshop and the prospect was discussed with CSO.

The three thematic sessions that comprised the workshop reflect SAS-114's focus on assessment and communication processes central to the evidence-based evaluation and improvement of intelligence processes and, by extension, intelligence product. The first session on Communicating Uncertainty, Source Reliability, and Information Credibility directed attention at a perennial challenge in intelligence communities: how to effectively communicate information about the uncertainty or confidence in intelligence assessments to decision-makers and other consumers of intelligence. It addressed questions such as, how can linguistic markers be effectively processed in ways that extract information from them in a reliable manner? How do intelligence organizations actually handle the communication of uncertainty and confidence by analysts? How do vague words of estimative probability change meaning across members in a NATO coalition who have diverse mother tongues? And how can we examine that and related issues experimentally within SAS-114?

The first session also included talks that focused on how information is communicated to (as opposed to from) analysts. That is, analysts receive information, which often is marked with "meta-informational" data. The information might be accompanied by indications of source reliability and information credibility. The integrity of that evaluation process for communicating such meta-information has not received much attention, and very early on in SAS-114's history (indeed, at its first meeting) the team agreed to that it would be wise to focus on this challenge. Thus, the first session also tackled questions such as, how does source reliability, information credibility, and classification level affect analysts' assessments of information accuracy? Do intelligence analysts reliably treat such marking when they use them to infer information accuracy? And how does NATO and national organizations advise intelligence personnel to assess and communicate such meta-information?

The second session on Structured Analytic Techniques in Intelligence Analysis addressed a central topic for twenty-first century intelligence studies and practice that nevertheless has received very little research attention or careful theoretical analysis (e.g., Refs. [12] and [15]). How can intelligence analysis, which is mainly a human-centric exercise in human judgment under conditions of uncertainty, be improved through structured techniques? How, for instance, can intelligence organizations effectively establish review or "devil's advocacy" processes that foster analytic integrity? How can computer-based techniques be leveraged to improve the computational effectiveness of intelligence analysis, and how could that be done without alienating the analyst to the point where the technique, even if effective, would not be used? And how effective are the structured techniques that are currently promulgated for use − those such as Heuer's [6] Analysis of Competing Hypotheses technique, which is routinely taught in intelligence analyst courses. These were among the daunting questions that speakers in the second session addressed and which the participants discussed.

The third session on Assessment of Information Quality and Risk was arguably the most diverse of the three sessions that comprised the workshop. Its presenters covered topics that raised questions such as: What is the true nature of information and how should we conceive of it? Is information a property of objects or of minds (or mind-object interactions)? How could the value of information as indicators best be assessed? And to what extent do formal models of information utility diverge? The session also focused on experimental work that examined how operators search for information, utilize it, change their beliefs on the basis of it, and ultimately decide based on their information search and evaluation processes. And last but not least, and ending on an eminently practical note, participants were carefully walked through core processes of a real-time capability for cyber security risk prioritization.

Beyond the presentations summarized in the report, there was a great deal of stimulating and productive discussion. Much of the discussion, moreover, focused on how the participants could effectively collaborate on research and best practice going forward. Since the workshop, there have been multiple teleconferences to explore collaborative work under SAS-114, and in those meetings further progress has been made. In short, the Workshop achieved its core objectives: it expanded the SAS-114 network, it allowed for an exchange of ideas and information, and it developed SAS-114's research agenda.

2.0 COMMUNICATING UNCERTAINTY, SOURCE RELIABILITY, AND INFORMATION CREDIBILITY

The substantive session on Day 1 of the Workshop focused on the topic of communicating uncertainty, source reliability, and information credibility. The last presentation described in this session, co-authored by Mandel and colleagues, was in fact delivered on Day 3. As well, it should be noted that five presentations given in this first session were not submitted for release in this report and are therefore not captured in this report.

Kicking off the session with a presentation entitled, **"**I think it is possible it might be so…Using Lexical Clues to Generate Evidence Weights" (see Annex A) Kellyn Rein, a DEU member of SAS-114, explained that an increasingly large amount of information is being automatically extracted from text using text analytics and computer linguistic algorithms. Much of this information is extracted and then stored in databases or ontologies for later use. However, as Rein noted, this can be problematic, as humans do not always communicate facts: they speculate, relate hearsay (not always accurately), and offer opinions. Thus, information extracted from text should be given some sort of credibility weighting, letting an analyst know how much credence should be put on that information. However, the high volume of textual information being produced today means that it is nearly impossible to have a human being evaluate extracted information for veracity. Therefore, methods for automating this process are needed.

Rein stated that when humans communicate they deliver information on two levels: the first and obvious level is the content information: "It is raining". The second level contains clues about the origin of that content (e.g., hearsay, assumption, inference) and how strongly the speaker holds to the veracity of the content. For example, the speaker may say "Mary told me it is raining", which indicates hearsay, or "It is quite possible that it is raining", which reflects the speaker's belief that it might be raining, without giving an indication as to where this came from. A third variation would be "I doubt that it is raining" which indicates the speaker considers it improbable that it is raining. In all three cases, the speaker indicates that there is uncertainty about the informational content of the utterance. Not all communications carry such markers, but when they do, it is possible to use such indicators to calculate an initial credibility value. By assigning values to various indicators of uncertainty such as “probably”, “possibly”, “doubtful” and scanning the immediate environment of the extracted information for these markers, analysts could automatically assign evidential weights. Linguists have performed a number of studies in which participants were requested to assign values to various such expressions. From study to study, the numerical values acquired varied, indicating that there is no “universal” value for, say, “probably”. However, there is a clear tendency to a fairly universal relative ordering of expressions, which can be exploited.

Rein noted, however, that humans very often include multiple indicators (e.g., “I think it’s possible it might be”), and use adverbs such as “very”, “quite”, “somewhat” to strengthen (called “boosters”) or weaken (called “down-toners”) the uncertainty expression, as well as sometimes negating the expression (“not possible”, “not unlikely”), shifting its meaning to its (near) opposite. Rein discussed an algorithm, which dealt with chaining multiple lexical elements in order to achieve a ranking of various, complex combinations of lexical elements in order to generate an initial evidential weight for extracted information.

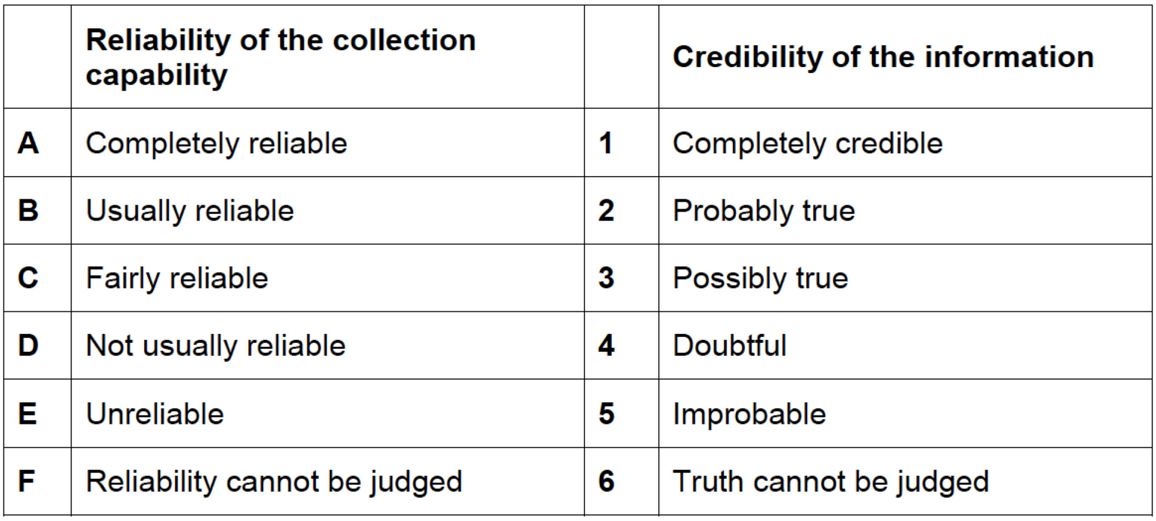
The next speaker, Christina Clarke, a GBR affiliate of SAS-114, provided an update on the UK defence intelligence approach to communicating uncertainty. Clarke outlined the status of projects focussed on uncertainty and the effect of recent organisational changes on them. She explained that Futures and Analytical Methods have a remit to improve analysis within Defence Intelligence through the provision of guidance and direct support to analysts and by determining or influencing organisational best practice. Since the last panel meeting, little has changed in guidance regarding uncertainty, in that analysts are still required to use the Uncertainty Yardstick when writing reports (a description of the Uncertainty Yardstick can be found in Ref. [7]). A tool for evaluating confidence in an assessment is still in production due to significant staff turnover. However, recent organisational changes mean that Futures and Analytical Methods should have more capacity to focus on analytic tradecraft, including the important areas of handling risk and uncertainty. Clarke noted that the research that could and should underpin such practices makes SAS-114 increasingly valuable, as does the ability to share experience with similar teams in other participating nations. In return, she recommended, there should also be more opportunity for analysts to participate in research projects when appropriate.

In a highly interactive session, James Kajdasz, a USA member of SAS-114, proposed a novel multinational experiment for SAS-114 in his presentation entitled, “Interpretation of NATO Standards by Non-Native English Speakers” (see Annex B). As Kajdasz explained, intelligence analysts must often communicate the likelihood that an event will occur to decision makers. One way to communicate likelihood is with a numeric probability (80% chance X will occur). However, verbal expressions (“it is likely X will occur”) are generally preferred by analysts and used in intelligence products. Using verbal expressions of probability increases the opportunity for miscommunication between analysts and decision makers.

As Kajdasz stated, there is a very large number of ways to communicate uncertainty verbally, and many expressions are interpreted differently between persons. In an attempt to reduce the chances for miscommunication, intelligence agencies have begun publishing lists of standard verbal expressions to be used when communicating likelihood verbally. Allied Joint Doctrine for Intelligence Procedures [13] has a list of five expressions of probability ranging from “highly likely” to “highly unlikely.” However, these probability terms are often interpreted by NATO members whose first language is not English. The non-English equivalent for a given English expression may be more optimistic or pessimistic than the English phrase’s connotation. Therefore, non-native English speakers from NATO countries may vary systematically in their interpretation of these English standards as a result of their native language influencing their interpretation of the English phrase. Accordingly, Kajdasz proposed that SAS-114 recruit non-native English speakers to evaluate the uncertainty communicated with the five NATO expressions. In his talk, he outlined a preliminary design that would allow differences between native languages to be evaluated. This was followed by open discussion of ways to shape the experiment and to work collaboratively on it through SAS-114.

Anne-Laure Jousselme, with collaboration from Francesca de Rosa, both of whom are NATO STO members of SAS 114 from the Centre for Maritime Research and Experimentation (CMRE), updated SAS-114 members and other Workshop participants on progress towards one of the key work elements of SAS-114, namely, the collection and analysis of defence and security standards for assessing and communicating source reliability, information credibility, uncertainty, and confidence (see Annex C). The team has now concluded the collection effort and will be turning attention to producing an interim report that summarizes the collection phase. An important subsequent task will involve carefully reviewing all of the standards and processes collected and providing a critical analysis of the various approaches, as well as recommendations for best practice in the future. This analysis will be summarized in the final report of SAS-114, which is scheduled for submission in December 2018.

The final presentation in this session by David Mandel (SAS-114 Chair and CAN member) and colleagues was entitled, "SAS-114 Experiment Update: Effect of Source Reliability, Information Credibility, and Classification Level on Analysts' Uncertainty about Information Accuracy" (see Annex D). Mandel provided an update on the experiment, which is designed to assess how meta-informational markers put on information in the "pre-analytic" evaluation step of the processing stage of the intelligence cycle influences intelligence analysts' assessments of the accuracy of the information they receive. The three sources of meta-information (i.e., information about information) examined in the research include source reliability, information credibility, and classification level. The coding of source reliability and information credibility are key functions of the evaluation process, and AJP-2.1 [13] describes the scales to be used for that purpose (see Figure 1).



	Reliability of the collection capability		**Credibility of the information**
A	Completely reliable	**1**	Completely credible
B	Usually reliable	**2**	Probably true
C	Fairly reliable	**3**	Possibly true
D	Not usually reliable	**4**	Doubtful
E	Unreliable	**5**	Improbable
F	Reliability cannot be judged	**6**	Truth cannot be judged

Figure 1: Scales for Source Reliability and Information Credibility in Allied Intelligence Doctrine [13].

Mandel and colleagues presented 42 intelligence analysts and defence operators with a full factorial combination of the levels of these scales (6 x 6), crossed additionally with a two levels of classification (TOP SECRET or OFFICIAL). Thus, participants received 72 unique meta-informational summaries and, for each, they judged the probability that the information that these markers pertained to was accurate. Degree of accuracy was measured on a probability scale ranging from 0 to 100 (i.e., percent chances). They were also retested on 10 of the 72 cases, and all 82 cases were presented randomly and on a trial-by-trial basis.

Consistent with other research showing that non-experts are influenced by the classification level or "secrecy" of the information [17], the findings showed that experts in the present experiment judged information to be more likely to be accurate if the information was described as Top Secret than if it was described as Official. As well, the findings showed that the reliability of experts' judgments decreased as the incongruence between the two scales increased. This was true both of intra-analyst reliability (i.e., test-retest consistency) and of inter-analyst reliability (i.e., the variability between participants). These findings are consistent with earlier results [1], [16] showing that users of these scales infer or otherwise assign scale levels that are highly congruent, if not perfectly matched, in terms of their ordinal value (i.e., A1, B2, C2,…, F6). Not only do experts tend to assign congruent values, the present findings show that experts also are uncertain what to infer from pairs of scale values that are not closely coupled. The SAS-114 experiment on meta-informational markers provides an evidence-based approach to assessing the effectiveness of a key piece of intelligence doctrine. As Mandel concluded, future research should examine alternative methods for encoding and combining meta-information for intelligence analysts.

3.0 STRUCTURED ANALYTIC TECHNIQUES IN INTELLIGENCE ANALYSIS

The first thematic session on Day 2 of the Workshop examined the topic of structured analytic techniques in intelligence analysis. Huib van de Meeberg and Alex Claver, both SAS-114 affiliates from NLD, kicked off the session with a presentation entitled, "Devil's Advocacy and Quality Assurance" (see Annex E). As they noted, the controversy regarding Iraq and its (non-existing) weapons of mass destruction called into question the validity of intelligence assessments and the effectiveness of existing quality control procedures within the intelligence services. The Dutch Ministry of Defence responded to the perceived need for better quality assurance by setting up a Devil's Advocate's office (DA) within the Dutch Military Intelligence and Security Service (NLD DISS). The Dutch DA has functioned since 2008 as a small organisational element, outside of the formal production process with unlimited access to all available information. To ensure the absolute independence of this unit it has been placed outside of the organisational hierarchy and is accountable to the Director only. As van de Meeberg and Claver explained, in this nearly a decade of activities the DA has seen two distinctive phases:

- **Phase 1: Performance Assessment: Focus on the "How" (2008 – 2011)**
  - Assessment and advice on quality improvement of intelligence products, processes (e.g., collection) and analytical skills to prevent and counteract tunnel vision, groupthink, etc.
  - Main tools were critical DA-surveys, product reviews (according to a fixed format and transparent quality criteria), the organisation of training courses and seminars as well as the establishment of a certified Master program within the Defence Academy ("improve the analyst").
  - These (still ongoing) activities are executed ex-post in order not to interfere with current production and intelligence research assignments. The focus is always to link the DA assessments with advice how to realise improvements in the future.
- **Phase 2: Performance Assessment: Focus on the "What and Why" (2012 – 2017)**
  - Do we do the right things and are we doing them effectively? To answer these questions, DA-activities include dealing with scarcity of resources through a system of qualitative 'Weighing and Prioritising' (W&P). W&P is supported by a quantification of available resource data and intelligence requirements and by assessments of the relevance of collections assets.
  - Customer feedback research: do the products and services meet the jointly set requirements?
  - Assessment of terminology and time indicators used in the analytical products and of the predictive value of analytical estimates (calibrated feedback).

In terms of results, van de Meeberg and Claver noted that, over the years, the DA has become widely accepted within NLD DISS, and quality improvements as well as greater sensitivity within NLD DISS to customers' interests were achieved. However, the effectiveness of a top-down approach has its limits. Progress remains fragile and does require continued 'investment' on all levels. Support for DA/QA by (highest) line management is essential.

Next, Floris Bex, also a NLD affiliate of SAS-114, discussed how computer-based computational systems could be used to augment the quality of intelligence production in a talk entitled, "Computational Scenarios and Arguments: An AI Approach to Structured Analytic Techniques" (see Annex F). Bex explained that despite the rise of "Big Data" and ever better algorithms to search for specific facts in large amounts of data, intelligence analysis is still very much driven by human thinking and decision making. Whilst computers are very good at, for example, sifting through huge datasets or computing the effect of new evidence on a large number of hypotheses, humans are needed to creatively construct new hypotheses, construct search queries and interpret the evidence in light of these hypotheses. Combining smart software and human analytic skills therefore seems like a good idea.

However, as Bex pointed out, the problem is that intelligence analysts work with natural language text, or semi-structured arguments and scenarios such as in Analysis of Competing Hypotheses (ACH) [6]. Computers, on the other hand, can only understand structured, mathematical models. If we want to build useful decision support software for analysts, we need principled ways of going from text to mathematical models, and back again (see Figure 2).

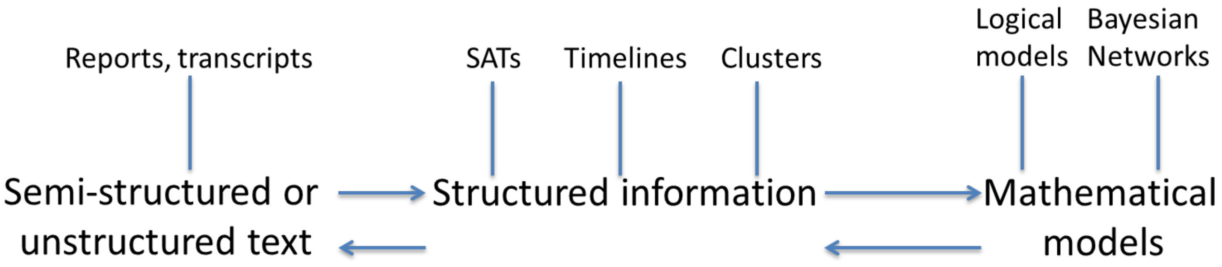


Figure 2: From Text to Mathematical Models and Back Again.

In these translations, we need to consider linguistic aspects (how do people express and interpret scenarios, probabilities, arguments in text?), mathematical aspects (what are the relations between qualitative and quantitative models?), design aspects (what are the goals of the system?), and psychological aspects (how do/should people reason?).

Bex provided a few examples of how human analysts can interface with logical and Bayesian models using natural language text or slightly more structured analytic techniques. The first example concerns a system that automatically parses and structures textual stories about a trade fraud case. The system subsequently checks any new story given the previous stories (e.g., is the new story compatible with what we know? is the new story complete?). The system then asks textual questions in order to further build a case. In his second example, Bex proposed a way to derive constraints for a Bayesian Network from structured arguments. Argumentation and Bayesian networks can both be considered decision support techniques, but are typically used by experts with different backgrounds. Bayesian network experts have the mathematical skills to understand and construct such networks, but lack expertise in the application domain, such as intelligence analysis, whereas analysts are typically more comfortable with argumentation approaches that are similar to familiar structured analytic techniques. As Bex explained, the proposed method allows us to check Bayesian networks given arguments constructed for the same problem, and also allows for transforming arguments into a Bayesian network structure, thereby facilitating Bayesian network construction.

The final talk in the session by Mandeep Dhami (a GBR member of SAS-114) and colleagues presented a research update entitled, "Report on SAS-114 Experiment on Analysis of Competing Hypotheses" (see Annex G). As Dhami explained, the Analysis of Competing Hypotheses (ACH) [6] is a structured analytic technique used in the intelligence community to help analysts identify alternative hypotheses and link evidence to hypotheses. ACH is meant to help analysts avoid confirmation bias (i.e., searching for evidence to support a favoured hypothesis or giving more weight to such evidence). ACH consists of eight steps as follows: Identify hypotheses, list significant evidence, create the ACH matrix, revise the ACH matrix, draw tentative conclusions, perform sensitivity analysis, report conclusions, and identify indicators for future observation.

Despite the popularity of ACH, there is a dearth of empirical research on ACH. Lehner *et al.* [11] found some evidence that ACH reduces confirmation bias in non-analysts but no evidence that it reduces bias in experienced analysts. Kretz *et al*. [9], [10] found that ACH did not do that much better than other techniques in, for example, generating more hypotheses. Finally, Trent *et al.* [18] were unable to evaluate the impact of ACH on analysis as, despite being trained on the technique, groups resisted using it.

The main aims of the present research were to examine how well analysts trained to use ACH can actually apply it in practice, and to explore the hypothesis-testing strategies that analysts' intuitively use. Dhami *et al.* used a between-subjects experimental design. Fifty intelligence analysts were randomly assigned to either be trained in using ACH (experimental group) or not (control group). They were each asked to perform an analytic task comprising four specific hypotheses and two general hypotheses. There were 12 pieces of evidence. The probability of occurrence of each piece of evidence was also provided, as was base-rate information.

The authors analysed the process that analysts in each group used to complete the task. It was found that both groups had a similar, but somewhat naïve understanding of the task, in terms of identifying the specific and general hypotheses and identifying the relevant pieces of evidence. For instance, only around two-thirds of the ACH group and also the control group identified all 12 pieces of evidence, and few focused on testing the two general hypotheses. Interestingly, whereas all of the ACH group represented the task information in a matrix as they were trained to do, the majority of the control group also did.

In terms of the reasoning strategies used, it was found that the control group of analysts that did not use ACH were more likely to use base-rate information than the ACH group. Dhami and colleagues also found that most analysts in both groups used an information integration strategy that gave weight to both the consistent and inconsistent evidence. This suggests that analysts who did not use ACH were not particularly prone to confirmation bias, and also that those who used ACH were resistant to the information integration method advocated in the technique, which only attends to disconfirmatory information. Interestingly, while all members of the control group drew a final conclusion that was compatible with their reasoning process, only around two-thirds of the ACH group drew a conclusion that was compatible with the reasoning depicted in their matrix. Finally, the control group demonstrated greater within-individual consistency of their hypothesis testing process compared to the ACH group. The experiment underscores the need for careful scientific verification of techniques, standards, or processes meant to improve human performance in defence and security environments (e.g., Ref. [4]).

4.0 ASSESSMENT OF INFORMATION QUALITY AND RISK

The afternoon of Day 2 and morning of Day 3 of the Workshop was devoted to the topic of information quality and its assessment. Jonas Clausen Mork, a SWE member of SAS-114, began the session with a presentation entitled, "Information Gain and Approaching True Belief" (see Annex H). As he explained, in many analyses of information, information bearers – or *infons* – are thought to be things like books, electromagnetic waves or sentences in natural language. While intuitively appealing, this can be difficult to combine with the idea the information carried by infons actually *inform*, in the sense of altering the belief state of a recipient. As infons also affect the beliefs of different recipients differently, assigning a fixed information value to a particular infon requires choosing some, possibly idealized, set of recipients with respect to which the infon has the "right" information value.

Another view, which Clausen Mork argues for, is that information values ought to be attached not to things like physical objects, signals or sentences, but to changes in belief. If something is to be called "information", certain properties appear reasonable, which in turn restrict the set of reasonable measures of information gain and loss. Two such measures were introduced in the presentation. One is "objective", increasing and decreasing as beliefs move closer and further away from true belief, or belief in the truth. For an observer for whom the truth is not known, the quantity of objective information gained or lost in a given belief change will also be unknown. The second measure is "subjective", allowing for introspective and inter-subjective assessment of gains and losses of information, even when the truth remains unknown.

While abstract, as Clausen Mork points out, this conception of information gain and loss can be connected to forecasting and the use of so-called *scoring rules*. As an example, the logarithmic scoring rule – an alternative to the widely used Brier score - corresponds to a penalty for the forecaster's subjective information gain as seen from the viewpoint of the scoring authority.

The next presenter, Jonathan Nelson, a DEU affiliate of SAS-114, delivered a talk entitled, "Optimal Experimental Design Theory, Asymmetric Cost Structures, and the Value of Information" (see Annex I). Nelson explained how Optimal Experimental Design (OED) models from statistics and philosophy of science could potentially be applied to evaluate the value of information on intelligence analysis tasks. For instance, when deciding which piece of information to acquire, to learn what group someone is a member of, one could choose information so as to maximize expected information gain. Frequently, people have thought of information gain as expected reduction in Shannon entropy in the probabilities (e.g., of each group that a person might be a member of). However, the Sharma-Mittal framework generalizes Shannon entropy, Hartley entropy, the Tsallis and Arimoto and Renyi families of entropy measures, and other ideas in a single two-parameter framework.

As Nelson describes, psychological research suggests that expected reduction of a moderate-order Arimoto entropy, or a similar entropy function, gives a good description of human data on many tasks, and has generally helpful mathematical properties. Which precise entropy measure best captures the psychology of uncertainty is not known, but the Sharma-Mittal framework helps us formulate questions for future experimental research with people.

Other important issues that Nelson and his colleagues have raised include how to incorporate asymmetric decision cost structures (e.g., if some mistakes are more harmful than others), finding the most helpful graphical or numeric formats for presenting probabilistic information to people, and figuring out how to convey the ways in which multiple pieces of information (e.g., SIGINT, and multiple human informants) should be combined. Nelson concludes that these issues can help suggest important mathematical and behavioral science research, in collaboration with intelligence community stakeholders.

In a related vein, Mark Timms, a CAN affiliate of SAS-114, discussed measures of information usefulness in target classification (see Annex J). As Timms explained, the scarcity of intelligence collection resources (e.g., Ref. [5]), and the inherent risks associated with the collection of sensitive information in support of foreign policy implementation, demand that information value be accurately identified in the earliest stages of operational planning. Setting out the practical context for work on information utility models, Timms points out that with a given mission defined, a commander's *critical information requirements* evolve into *priority intelligence requirements*, which ultimately guide which intelligence resources are dedicated to collecting specific pieces of information intended to enable command decision making [2]. In essence, once a commander (or any other decision-making stakeholder) understands what their superiors expect them to accomplish, they will inevitably shift focus to prioritizing the information they require in order to optimize mission-specific decision-making. Ultimately, information requirements are identified through a multi-stage process through which groups of people collaboratively debate the validity of their personal opinions formulated through their subjective understanding of the tasks at hand. The success of this process is entirely dependent on the abilities of the individual planner (or planning group), as well as the diversity, size, and breadth of experience (among other items) of the members involved [3]. Intuitively, the outputs of this process are highly vulnerable to human error. As Timms highlighted, evolving information utility research could have immediate positive impact in this domain.

Timms reviewed six quantitative models of assessing information utility, which were described in Nelson [14] − specifically, Bayesian Diagnosticity, Log Diagnosticity, Information Gain, Kullback-Leibler Distance, Probability Gain (error minimization), and Impact. Nelson grouped the models into two distinct approaches to evidence acquisition: falsification (disproving a given hypothesis – correct vs. incorrect) and differentiation (reducing hypothetical uncertainty through clustered evidence acquisition – most or least plausible). As Timms explained, Nelson found that the mathematical structure of some of the more established and popular models (Bayesian Diagnosticity and Log Diagnosticity) appeared to favour unique pieces of evidence that disprove a given hypothesis no matter how rare that evidence might be in a given framework, whereas differentiation models more accurately identified individual (and clustered) pieces of evidence that assisted with the gradual reduction of hypothetical uncertainty. As Timms noted, he and other SAS-114 members are beginning to explore the application of such quantitative models of information utility to optimizing intelligence collection efforts.

Turning from formal models to experimental methods, Anne-Laure Jousselme, in her presentation entitled “Risk Game: Impact of Information Quality on Decision Making” (see Annex K), described a task that could be used to study information utilization and the influence of information sources on beliefs about alternative hypotheses pertinent to decision-making. The Risk Game is a general methodology developed at CMRE to elicit experts’ knowledge and know-how, including their ability to deal with information of different nature (from sensors to human witnesses), to consider the information quality (including source quality) and to reason about concurrent events [8]. It is a hypothetical-scenario-based technique aimed at capturing data expressing human reasoning features while performing a specific task of maritime situation assessment (although the task could be adapted for research in other environments).

The main focus of the Risk Game is the study of the impact of information quality on the ability of human operators to assess threat and make decisions. A new methodology is presented where information is abstracted away by cards, and the quality is randomly selected by dice roll. The game has been tested during the Table Top Exercise (TTX) for harbour protection held at CMRE in November 2014. The game has been played by 32 experts, most of them OF-3 and above of the maritime domain, from 9 NATO nations. As Jousselme noted, the preliminary results obtained are promising and allowed her and her colleagues to identify future research challenges. Among key results, she found that the players' perceived relevance of information may differ from the effective relevance (i.e., deviation from optimal use of information), that a high amount of false information increases the uncertainty of the player before decision and may lead to wrong decisions, and that the context can have a high impact on the decision taken. An aim of future work with the Risk Game will be to formalize its aspects in order to better support the analysis of players’ reasoning profiles in subsequent experiments.

The final presentation of the session, delivered by Greg Weaver, a USA member of SAS-114, described Army Research Laboratory’s Information Systems Continuous Monitoring (ISCM) capability for cyber security risk prioritization (see Annex L). As Weaver explained, ISCM is an integrated big-data capability that provides situational awareness down to the asset level as well as cyber security risk prioritization and categorization for multiple enclaves, as each asset object is a fusion of data gathered from reports generated via the endpoint security, vulnerability assessments and intrusion detection system data flows. These data sources were chosen because they are ubiquitous across the US Department of Defense, but they could be easily swapped with other data sources if the attributes of interest were available for extraction.

As Weaver noted, the ISCM big-data capability supports the overarching goal in efforts to conduct research and development to deploy an operational framework to:

1) Enhance cyber situational awareness – The ability to ingest, aggregate, correlate and enrich cyber data from a variety of sources and provide an interface or dashboard view that enables commanders and missions owners to make higher-confidence decisions and prioritize cyber security risks and responses.

2) Support continuous monitoring – The ability to transform the historically static security control assessment and authorization process into an integral part of a dynamic enterprise-wide risk management process. Providing the Army with an ongoing, near real-time, cyber defence awareness and asset assessment capability.

3) Enable technical transfer – The ability to be packaged and transitioned to other organizations with a similar cyber security mission and data sets. In particular, it is important that ISCM be transferable with minimal software refactoring and systems reengineering.

4) Provide a scalable architecture – The ability to scale quickly be augmented with minimal impact to uptime and support the storage and processing of large data sets at the TB/PB scale.

5) Enable low latency queries – The ability to provide rapid responses to simple and compound queries from both end users and statistic/analytic processes (query focused).

5.0 REFERENCES

[1] Baker, J.D., McKendry, J.M. and Mace, D.J. (1968). *Certitude judgments in an operational environment*. ARI Technical Research Note 200 (AD 681 232). Arlington, VA: US Army Research Institute for the Behavioral and Social Sciences.

[2] Chief of Defence Staff. (2002). *CF Operational Planning Process.* B-GJ-005-500/FP-000. Department of National Defence: Ottawa, ON.

[3] Dhami, M.K., Belton, I. and Careless, K. (2015). *A review of analytic techniques*. Report prepared for HM Government, UK. Available from first author.

[4] Dhami, M.K., Mandel, D.R., Mellers, B.A. and Tetlock, P.E. (2015). Improving intelligence analysis with decision science. *Perspectives on Psychological Science, 106*(6), 753-757.

[5] Folker, R.D., Jr. (2000). *Intelligence analysis in theater joint intelligence centers: An experiment in applying structured methods*. Joint military intelligence College, Washington DC Center for strategic Intelligence Research.

[6] Heuer, R.J., Jr. (1999). Psychology of Intelligence Analysis. Washington, D.C.: Center for the Study of Intelligence, Central Intelligence Agency.

[7] Ho, E., Budescu, D.V., Dhami, M.K. and Mandel, D.R. (2015). Improving the communication of uncertainty in climate science and intelligence analysis. *Behavioral Science and Policy, 1*(2), 43-55.

[8] Jousselme, A.-L., Pallotta, G. and Locke, J. (2015). *A Risk Game to measure the impact of information quality on human threat assessment and decision making*. NATO Technical Report [CMRE-FR-2015-009, NATO UNCLASSIFIED].

[9] Kretz, D.R. and Granderson, C.W. (2013). An interdisciplinary approach to studying and improving terrorism analysis. In *2013 IEEE International Conference on Intelligence and Security Informatics (ISI).* DOI: 10.1109/ISI.2013.6578808.

[10] Kretz, D.R., Simpson, B.J. and Graham, J. (2012). A game-based experimental protocol for identifying and overcoming judgment biases in forensic decision analysis. In *2012 IEEE Conference on Technologies for Homeland Security (HST)*. DOI: 10.1109/THS.2012.6459889.

[11] Lehner, P.E., Adelman, L., Cheikes, B.A. and Brown, J.J. (2008). Conformation bias in complex analyses. *IEEE Transactions on Systems, Man and Cybernetics – Part A: Systems and Humans, 38*, 584-592.

[12] Mandel, D.R. (2009). *Applied behavioural science in support of intelligence: Experiences in building a Canadian capability*. Commissioned Report to the Committee on Field Evaluation of Behavioral and Cognitive Sciences-Based Methods and Tools for Intelligence and Counter-intelligence, Division of Behavioral and Social Sciences and Education.

[13] NATO Standardization Office. (2016). *AJP-2.1, Edition B, Version 1: Allied Joint Doctrine for Intelligence Procedures*. Brussels, Belgium: author.

[14] Nelson, J.D. (2005). Finding useful questions: on Bayesian diagnosticity, probability, impact, and information gain. *Psychological review*, *112*(4), 979.

[15] Pool, R. (2010). *Field evaluation in the intelligence and counterintelligence context: Workshop summary*. Washington, DC: National Academies Press.

[16] Samet, M.G. (1975). *Subjective interpretation of reliability and accuracy scales for evaluating military intelligence*. ARI Technical Paper 260. Arlington, VA: US Army Research Institute for the Behavioral and Social Sciences.

[17] Travers, M., Van Boven, L. and Judd, C. (2014). The secrecy heuristic: Inferring quality from secrecy in foreign policy contexts. *Political Psychology, 35*(1), 97-111.

[18] Trent, S., Voshell, M. and Patterson, E. (2007). Team cognition in intelligence analysis. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 51,* 308-312.