CREST Centre for Research and Evidence on Security Threats

Taking Decisions About Information Value

FULL REPORT

SEPTEMBER 2020

Ashraf Labib
Salem Chakhar
Lorraine Hope
John Shimell
Mark Malinowski
Tom Baldwin

SEPTEMBER 2020

Taking Decisions about Information Value

FULL REPORT

Ashraf Labib, University of Portsmouth

Salem Chakhar, University of Portsmouth

Lorraine Hope, University of Portsmouth

John Shimell, Polaris Consulting Limited

Mark Malinowski, Polaris Consulting Limited

Tom Baldwin, NATO Allied Command Transformation (ACT)

This is the Full Report from the *Taking Decisions about Information Value* project, commissioned by CREST. The project aims to enable analysts to achieve better judgements about the value of elicited information from intelligence reports.

www.crestresearch.ac.uk/projects/taking-decisions-information-value/

This research was funded by the Centre for Research and Evidence on Security Threats – an independent Centre commissioned by the Economic and Social Research Council (ESRC Award: ES/N009614/1) and which is funded in part by the UK Security and intelligence agencies and Home Office.

www.crestresearch.ac.uk

TABLE OF CONTENTS

PART I: EXECUTIVE SUMMARY

This report comprises the findings of CREST funded research into project into making decisions about information value. It addresses an important challenge for intelligence analysts. Intelligence analysts are typically required to process large volumes of data in a timely manner in order to extract useful information and detect potential security threats.

This relies on consistent judgements by the analyst in order to efficiently process the data and effectively identify useful information. Research and historical evidence have shown that analysts' judgements are often inconsistent due to the mass of data, the variation in types and nature of intelligence information and the time pressures the analyst is operating within. Consequently, intelligence analysts will often take decisions that deviate significantly from those of their peers, from their own prior decisions, and from training rules that they themselves claim to follow. Such inconsistency is mainly due to two types of errors; noise and bias, which complicate the intelligence analysis process and can result in key pieces of data being misclassified or overlooked with potential security threat implications.

The proposed project aims to develop, train and evaluate an innovative analytic approach to address these errors and enable analysts to achieve better judgements about the value of elicited information from intelligence reports. The innovation is in the embedding a machine learning method called the Dominance-based Rough Set Approach (DRSA) algorithm within a tool that enables an intelligence analyst's interests and behaviour to be captured. This is designed to evaluate the consistency of analysts' judgements at individual and group levels, as well as identifying key factors or biases which influence an analyst's decision making.

The findings were used to inform analyst training and as a decision aide within the tool to ensure more robust judgements are made.

Specifically, the project aims to address the following research questions:

1. By how much does individual analyst bias affect the quality of the decisions?
2. Will incorporation of group decision support, as opposed to individual support, improve the quality of decisions?
3. Do additional facilities of feedback in consistency and sensitivity analysis provide a support for better decision taking?

The DRSA tool is based on a prototype that was originally developed under a Ministry of Defence (MOD) innovation competition in 2016 and the further development of the tool in the course of the propose project comprised three layers.

The first layer uses a classic DRSA machine learning approach to deduce relevant analyst insights from a subset of intelligence reports that can be applied to predict relevance of unseen intelligence reports.

The second layer applies an innovative aggregation procedure (Chakhar, Ishizaka, Labib, & Saad, 2016) which incorporates a group decision taking facility into the tool to make it more relevant to an actual intelligence analysis team.

The third layer is devoted to validation and sensitivity analysis. This layered structure resulted in the development of a powerful decision tool that can be dynamically updated to monitor and capture analyst judgements to support training and provide real-life support to intelligence analysts. The existing DRSA

prototype tool, which was prior to the current project, covered only the first layer.

The project was conducted over three phases: Phase 1 works with relevant stakeholders to capture data, and to design the tool to meet their specific needs. Phase 2 is concerned with the development and testing of the tool and Phase 3 consists of an experimentation phase via systematic testing with relevant stakeholders during training workshops and experiments.

The project was conducted by a multidisciplinary research team from academia and industry. The University of Portsmouth provided experts in root cause decision analysis and Operational Research (OR) techniques (Professor Labib), information elicitation (Professor Hope), multi-criteria decision analysis and DRSA (Dr. Chakhar) and criminal intelligence (Dr. James). Polaris Consulting provided expertise in OR, experimental design and software development within the MOD intelligence domain.

The significance of the proposed approach is that it provides a tool that is user friendly yet capable of providing sophisticated analytics that can substantially improve decision taking, minimize risks, and unearth valuable insights that may otherwise have been lost. Since the team includes experts from security in both police and military contexts, one of the perceived benefits of exploitation is that there will be cross-fertilisation between police intelligence and the military domains.

Perceived benefit for security analysts is to continuously adapt and learn from previous mistakes using an approach that systematically improves accuracy and consensus in judgements in a timely manner. Figure 1.1 shows a typical scenario of analyst scope. Figure

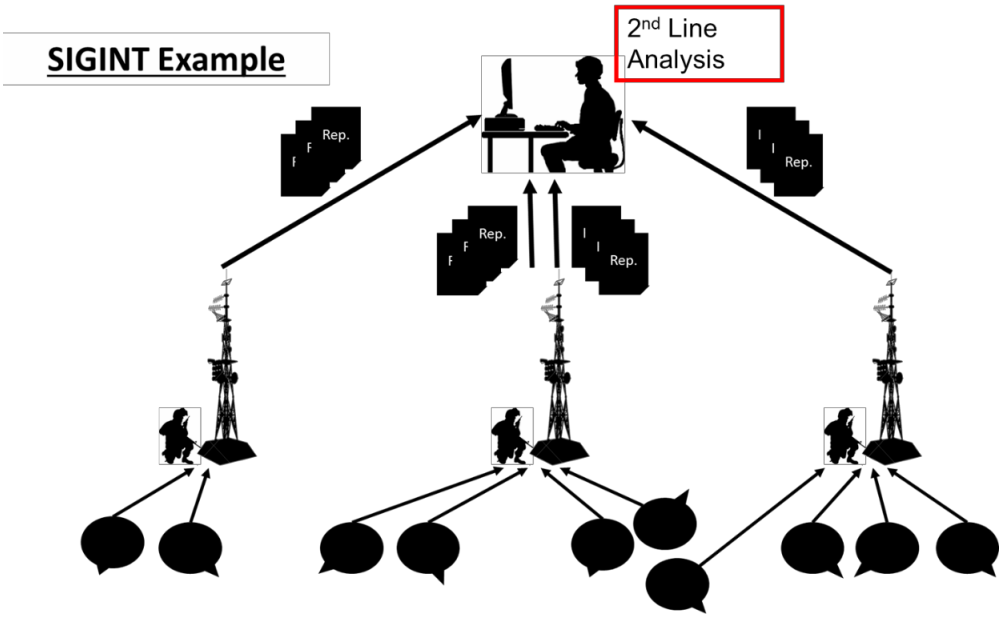

Figure 1.1 Scenario analyst scope

Figure 1.2 Principle of DRSA

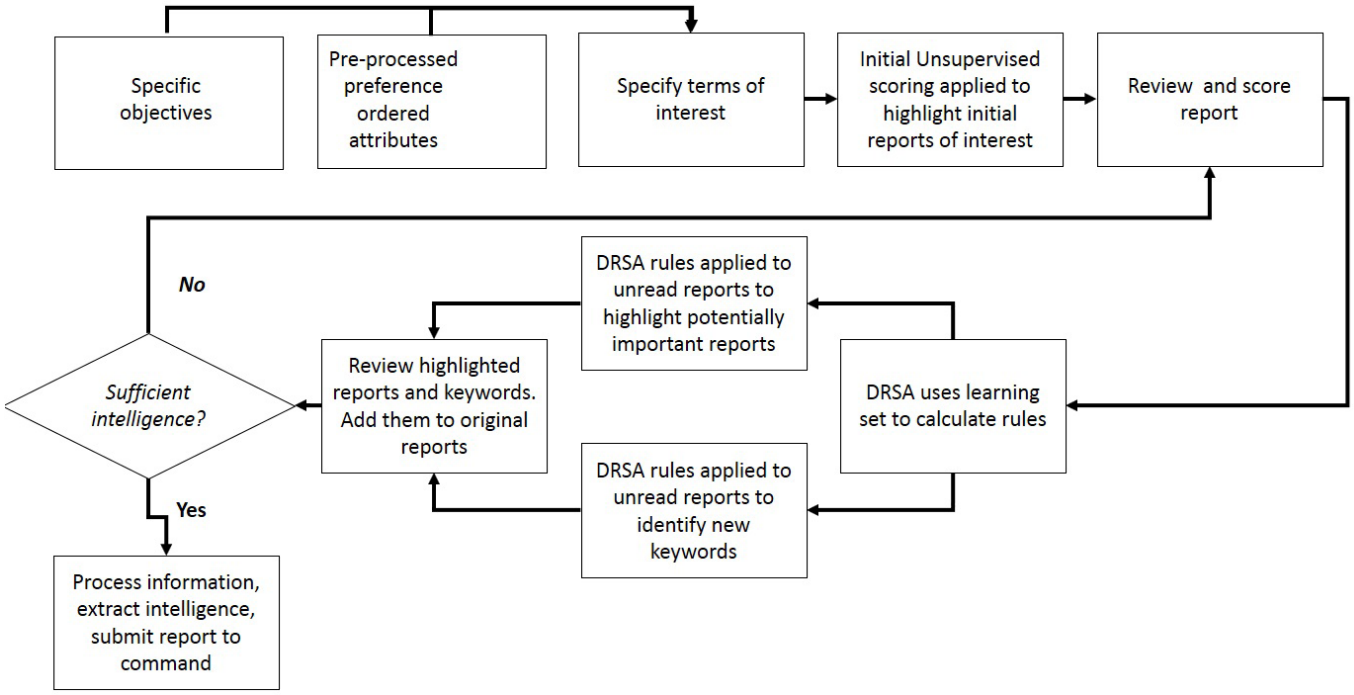

Figure 1.3 Flowchart of typical intelligence reports assessment exercise for single analysts

1.2 shows the basic principle of DRSA method. And figure 1.3 shows the basic idea in the form of a typical intelligence reports assessment exercise for intelligence analysts.

The report consists of three parts.

Part II is the core of the project and contains theoretical background, the tool design and its development, as well as validation and assessment of results. This part is also the base of an intended journal paper to be submitted.

Part III is related to the design specifications of the software and planning for the testing sessions that we have conducted. Finally, Part IV summarizes disseminations of results

PART II: ANALYSIS OF NOISE AND BIAS ERRORS

Analysis of noise and bias errors in intelligence analysts' judgements (Software Implementation and Contribution to Theory and Practice)

Ashraf Labib, Portsmouth Business School, University of Portsmouth, UK

Salem Chakhar, Portsmouth Business School, University of Portsmouth, UK

Lorraine Hope, Department of Psychology, University of Portsmouth, UK

John Shimell, Polaris Consulting Limited, TP Group plc, Cody Technology Park, Farnborough, UK

Mark Malinowski[3], Polaris Consulting Limited, TP Group plc, Cody Technology Park, Farnborough, UK

Tom Baldwin, NATO Allied Command Transformation (ACT) in Norfolk, Virginia, USA.

ABSTRACT

Intelligence analysts are typically required to process large volumes of data in a timely manner in order to extract useful and, ideally, actionable information and detect potential security threats.

Research and historical evidence have shown that analysts' judgements are often inconsistent due to the mass of data received, the variation in types and nature of intelligence information and the time pressures the analyst is operating within.

Such inconsistency is mainly due to two types of decision-making errors: noise and bias. These errors complicate the intelligence analysis process and can result in key pieces of information being misclassified or overlooked with potential security threat implications.

The team-oriented aspects of intelligence analysis process further complicate the situation. To identify and address these errors and enable analysts to achieve better judgements about the value of information contained in intelligence reports, we have designed, implemented and validated an innovative decision support platform.

This paper reports on a series of validation trials in which research participants took part in a simulated intelligence assessment task and were required to make decisions about in-coming intelligence information.

We sought to assess the performance of individual and team-oriented analysts with respect to effectiveness and efficiency by examining patterns of 'hits', 'misses' and 'false alarms'.

The main results of these simulations are: (1) the use of the developed tool (a) ameliorates slightly the effectiveness of individual intelligence analysis process and (b) improves considerably its efficiency; and (2) the incorporation of group decision support improves largely both the effectiveness and efficiency of intelligence analysis process.

1. INTRODUCTION

Modern intelligence analysis is complicated by the large volumes of incoming data that should be processed in a timely manner in order to extract useful, and ideally, actionable information and detect potential security threats. The analysis task relies on consistent judgements by the analyst in order to efficiently process the data and effectively identify useful information. Research and historical evidence (see e.g. DeRosa, 2004; Fischhoff & Chauvin, 2011; Renya et al, 2014) has shown that analysts' judgements are often inconsistent due to a number of factors including the sheer mass of data, the variation in types and nature of intelligence information and the time pressures the analyst is operating within.

Consequently, intelligence analysts will often take decisions that deviate significantly from those of their peers, from their own prior decisions, and from training or operational rules that they themselves claim to follow. Such inconsistency is mainly due to two types of errors: noise and bias. Here noise is defined as the variability of judgements or inconsistent decisions, whereas bias is defined as consistent diversion from the target (Adame, 2016; Hammond et al, 2006; Kahneman & Rosenfield, 2016). These errors complicate the intelligence analysis process and can result in key pieces of data being misclassified or overlooked with potential security threat implications (Heuer, 1999; Reyna et al, 2014)[1].

The situation is further complicated when considering team-based decision making, as opinions and unintentional (or intentional) biases within groups can lead to inconsistency and a lack of consensus. Due to the real-time working context, communicating pertinent information (e.g. some relevant reports with crucial information about a planned attack) identified by one member of the analysts team to other team members become difficult since these team members are overwhelmed by the incoming reports and have to pause to absorb other new data. Thus, team-based decision making, or analysis of input from several intelligence agents presents a challenge to modern intelligence analysis due to the inherent inconsistency that individual as well multiple opinions can lead to.

An innovative analytic approach to identify and address inconsistency and enable analysts to achieve better judgements about the value of elicited information from intelligence reports has been designed and implemented. The innovation is in the embedding of the Dominance-based Rough Set Approach (DRSA) (Greco et al, 2001) within a tool that enables an intelligence analyst's interests and behaviour to be captured. The developed tool supports two levels of analysis.

The first level uses the DRSA to deduce relevant analyst insights from a subset of intelligence reports that can be applied to predict relevance of unseen intelligence reports.

The second level applies an innovative aggregation procedure (Chakhar et al, 2016) which incorporates a group decision taking facility into the tool to make it more relevant to an actual intelligence analysis team.

In a study reported in Baldwin et al (2016)[2], a DRSA-based proof-of-principle tool was successfully demonstrated within an experiment that used volunteer intelligence analysts and example Signals Intelligence (SIGINT) data. The main findings showed that the DRSA tool made a sizeable difference by enabling analysts to identify a greater proportion of relevant

1 These two references mentioned about biases explicitly. Noise was also mentioned but implicitly, as they called it disagreement, or fragmented opinions etc.

2 This study concerns a project funded under the UK Ministry of Defence (MoD) Centre for Defence Enterprise (CDE) Autonomy and Big Data competition.

reports (by about 19%) and filter out irrelevant reports.

The DRSA-based tool offers significant potential compared to other multi-criteria and machine learning approaches as: it accepts any type of data including numbers and text, it is able to deal with incomplete/ missing data and it is able to detect and deal with inconsistency problems. In other words, inconsistency in DRSA is treated as natural, rather than an outlier case.

The main advantage of the DRSA tool compared to other methods is the simplicity of its approach in that it uses the learning set to derive understandable If-then rules, which can be used to analyse judgements and provide feedback with a low training burden. DRSA has also been shown to be more accurate than alternative classification methods including nearest neighbours, support vector machine, decision trees, multi-layer perceptron network, fuzzy classification and naïve Bayes approach (Chakhar et al, 2016; Hu et al, 2017).

To validate the developed tool a series of trials were designed and conducted by the authors, where the task is to identify and prioritise intelligence information pertaining to a critical target event and that involve security professionals. The paper looks in particular to assess the performance of individual and team-oriented analysts with respect to effectiveness and efficiency by examining patterns of 'hits' (high score given to important reports), 'misses' (low scores given to important reports) and 'false alarms' (high scores given to unimportant reports).

More specifically, this paper reports the results of the analysis of noise and bias errors in intelligence analysts' judgements using the developed tool as a decision support platform. The main results of the analysis are: (1) the use of the developed tool (a) ameliorates slightly the effectiveness of individual intelligence analysis process and (b) improves considerably its efficiency; and (2) the incorporation of group decision support improves largely both the effectiveness and efficiency of intelligence analysis process.

The rest of this paper is organized as follows. Section 2 provides the background. Section 3 introduces some theoretical aspects. Section 4 addresses tool design and development issues. Section 5 presents the tool validation and results analysis and discussion. Section 6 concludes the paper.

2. BACKGROUND

2.1 INCONSISTENCY IN INTELLIGENCE ANALYSIS

Intelligence analysts are typically required to process large volumes of data in a timely manner in order to extract useful information and detect potential security threats. This relies on consistent judgements (i.e. identical cases should be treated similarly, if not identically) by the analyst in order to efficiently process the data and effectively identify useful information. In most instances, the analyst will be trained to recognise typical threat characteristics when reviewing data.

For example, the analyst will review data to identify individual or group level suspicious activity or communications, which may be indicative of larger-scale criminal or terrorist activities. It is vital that analysts, often operating as part of team, are able to make consistent judgements about the value of this information, so it can be quickly extracted enabling appropriate security and counter-terrorism measures to be taken.

However, research and historical evidence (DeRosa, 2004; Fischhoff & Chauvin, 2011; Renya et al, 2014; Heuer & Randolph, 2015) have shown that analysts' judgements are often inconsistent due to the mass of data, the variation in types of data, the lack of evidence presented, and the time pressures the analyst is operating within. Such inconsistency complicates the intelligence analysis process and can result in key pieces of data being misclassified or overlooked. The potential impact of this is to reduce the effectiveness of security and counter-terrorism resources (e.g. mistakenly deployed against wrong target), damage their reputation (e.g. not identifying a terrorist threat) and, in the worst case, result in an increased likelihood of criminal and terrorist activities.

2.2 NOISE AND BIAS ERRORS

Intelligence analyst judgements or decisions reflect the fallibility of humans as rational decision-makers. Indeed, theory and research amply demonstrate the shortcomings of human decision processes, noting a myriad of biases and errors (Adame, 2016; Hammond, Keeney & Raiffa, 2006; Hills, 2016; Tversky & Kahneman, 1974). Since judgements are concerned with the selection and ranking of choices, a key problem is that in the absence of adequate feedback that is both immediate and clear, humans are also unreliable decision takers. Research suggests that the unavoidable consequence is that professionals (such as intelligence analysts) will often take decisions that deviate significantly from those of their peers, their own prior decisions, and from rules that they themselves claim to follow (Kahneman & Rosenfield, 2016; Sigurdsson, 2016). This inconsistency or errors within judgement is mainly due to two types of errors (Kahneman 2011; Kahneman & Rosenfield, 2016): noise and bias. Kahneman & Rosenfield (2016) coined the term 'noise' to describe the chance variability of judgements or inconsistent decisions. Another type of error in expert judgement is 'bias' which indicates consistent diversion from the target (Adame, 2016; Hammond et al, 2006;). Such bias leads to high variation within decision making and can be categorised into:

(i) social and cognitive biases, such as anchoring bias (when the estimation of a numerical value is based on an initial value – anchor – which is then insufficiently adjusted to provide the final answer) (Tversky & Kahneman, 1974);

(ii) status quo bias (when people ask to get paid more for an item they own than they are willing to pay for it when they do not own it; their disutility for losing is greater than their utility for gaining the same amount) (Kahneman et al, 1991) sunk-cost confirming-evidence (people consider sunk cost when making prospective decisions) (Arkes & Blumer, 1985); and

(iii) framing bias (when people characterise

the initial problem in a certain flawed way) (Tversky & Kahneman,1981).

For an excellent review of cognitive and motivational biases in decision risk analysis and means of overcoming them (debiasing), the reader is directed to the work of Montibeller & von Winterfeldt (2015). Here the authors define motivational biases as "those in which judgments are influenced by the desirability or undesirability of events, consequences". An example of a motivational bias is the deliberate attempt of experts to provide optimistic forecasts for a preferred action or outcome. Another example is the underestimation of the costs of a project to provide more competitive bids (Montibeller & von Winterfeldt, 2015). Research to date suggests that an appropriately designed algorithm, which corrects and calibrates inconsistent judgements, might be used to minimise these errors by improving the quality of decisions in a timely manner (Kahneman & Rosenfield, 2016).

In studying the effect of noise and bias errors on decision accuracy, Kahneman & Rosenfield (2016) distinguished four situations: (i) accurate decision (i.e. no noise, neither bias); (ii) noise when there is variability within the judgements of different decision makers (or from the same decision maker on the same data); (iii) bias when judgements are similar but not correct (consistently wrong); (iv) noise and bias when both (ii) and (iii) situations hold. Within the considered intelligence analysis context, noise errors hold when intelligence reports are judged differently by different intelligence agents. The level of noise will increase with the distance between the risk levels of the target reports as specified by the different agents. In the same context, bias errors hold when intelligence reports are wrongly judged by the different intelligence agents, but they are consistently wrong; in other words, there is consensus on the wrong interpretation or decision based on the available data. The noise and bias errors can jointly hold, leading to additional potential security threat implications.

2.3 MEASURING AND CORRECTING NOISE AND BIAS ERRORS

Noise and bias affect accuracy differently. Figure 1 illustrates graphically the effect that noise and bias can have on accuracy. In this figure, a Likert risk scale of five levels has been assumed where level 5 is the highest risk level. Decisions (designed by ✖ in Figure

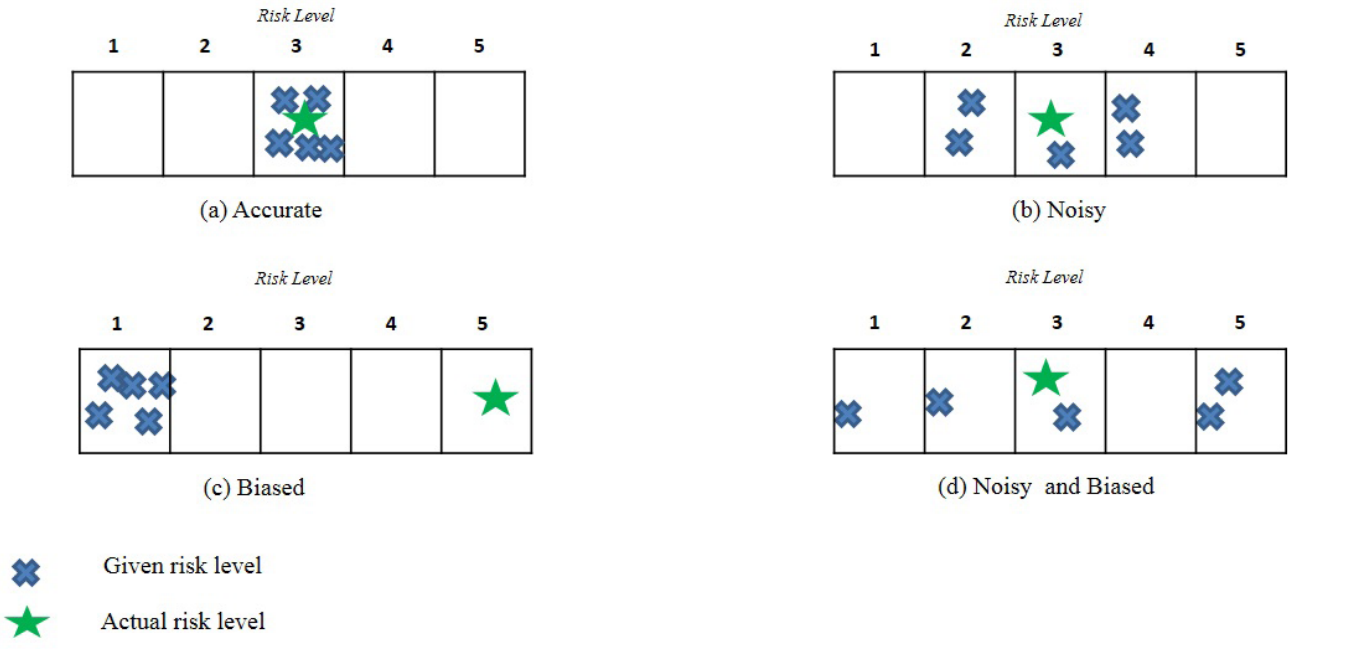

Figure 1 Noise and bias situations

1) in situation (a) are accurate since scores provided by different analysts are accurate since that are close to one another. The other three situations are inaccurate but in distinctive ways. In situation (b), decisions are noisy since the scores are widely scattered around the correct score (designed by ★ in Figure 1). In situation (c), decisions are biased since they miss the correct decision but cluster together. In situation (b), decisions are both noisy and biased. Note here (as in Kahneman & Rosenfield (2016)). that no hypothesis is made that the quality of judgment is measurable.

Noise and bias errors are measured and corrected differently. The noise error can be measured through dispersion statistics or in terms of the distance of scores to each other. It is important to note that there is no need to know the correct answer to measure noise (Kahneman & Rosenfield, 2016). One possible solution to correct noise error is to aggregate scores using simple rules such as average or median/mode. However, average rule is not working in the application considered in this paper because the risk scale is assumed ordinal. Scores can also be transformed into numerical ones (using some appropriate transformation methods) and then average and other similar aggregation rules can be used. The use of rank aggregation techniques is another possible solution. In all cases, correction can be handled automatically without the need to advice analysts.

The bias error can be measured in terms of the distance to the correct score. Here, we need to know the correct answer to measure bias (Kahneman & Rosenfield, 2016). The automatic correction of bias error is difficult and may complicate the situation (adding more bias or noise). A better solution is to ask analysts to revise their decisions.

2.4 STRUCTURED ANALYTIC TECHNIQUES

Within the Intelligence Community, professionals often rely on a set of Structured Analytic Techniques (SAT) for reducing noise and bias. The SAT is a set of mental modelling tools of externalizing, organizing and evaluating analytic thinking. They aim at providing a 'contrary' analysis in order to challenge the status quo and consensus view of analysts. In doing so, they encourage questioning existing equilibrium and status quo, promote imaginative thinking, and help to develop alternative outcomes, which can all support debiasing (Montibeller and Von Winterfeldt, 2015). In addition, the use of a structured process helps in depersonalizing arguments and hence can support objective group decision making. The proponents of SATs argue that they are grounded on the concept of System 1 (intuitive-based) and System 2 (analytic-based) thinking, where they mainly focus on System 2.

There are about twelve basic SAT that can be broadly categorized into diagnostic, contrarian, and imaginative. Examples of diagnostic techniques are key assumptions, and quality of information checks, and indicators and signposts of change. Such diagnostic techniques can address cognitive biases such as status quo and anchoring type of biases. Examples of contrarian techniques are analysis of competing hypothesis (ACH), devil's advocacy, and what-if analysis, and they tend to address status quo and confirmation types of biases. Finally, examples of imaginative biases are brainstorming, outside-in thinking, and alternative future analysis. Again, these SAT tend to address both status quo and confirmation types of biases.

SAT are considered a gold standard in intelligence analysis training programs, as they are intended to prevent and mitigate against the two main sources of errors in judgments; systematic biases, and random noise. The proponents of SAT claim that they improve judgment quality through debiasing, organize complex

evidence, and promote rigorous and transparent analysis (Heuer et al, 2010). The list of SATs is not static. For example, in its latest version SATs have expanded to 55 tools in eight categories (Heuer & Randolph, 2015).

It has been argued that operational research (OR) can contribute to 'connecting the dots', which is a fundamental challenge in intelligence analysis and is about selecting and assembling fragmented pieces of information to produce a pattern that can improve understanding of a potential threat (Fischhoff & Chauvin, 2011; Kaplan, 2011). The proposed solution in the next section outlines an OR technique that can help in this task.

3. THEORETICAL DEVELOPMENTS

3.1 LIMITATIONS OF STRUCTURED ANALYTIC TECHNIQUES

The benefits of using SAT in intelligence analysis is well recognized. However, it has been argued that SAT suffer from flaws in both conception and design (Cheng et al, 2018). In particular, they treat bipolar bias as unipolar, which may result in over-shooting; for example, transferring an under-confidence to over-confidence and vice versa. Moreover, since they focus on problem decomposition, in doing so they may also have a negative impact on reliability of assessment, which leads to increase of noise, the very type of error they intend to overcome in the first place. The same authors also argue that SAT have not been subjected to rigorous empirical assessment. In other words, it is not clear whether in their attempt to solve problems, they are creating more problems, or even just being ineffective. This is akin to the notion that having antibiotics as a drug to cure illness is good as a concept, but an over-dose can be toxic. This state of affairs in the intelligence domain can either cause wrong decisions being made, or cause a startled and paralyzed delayed process, which questions their value added to decision support.

In addressing such limitations of SAT, one of the main suggestions of Cheng et al (2018) is to establish more explicit set of rules for handling evidence. In the next section, we describe a proposed solution that relies on provision of explicit rules that are either 'certain' or 'possible' based on data related to evaluation of set of attributes.

3.2 PREFERENCE AND BEHAVIOUR LEARNING APPROACH

The project[3] considered in this paper aims to develop and test whether an innovative approach that provides feedback to security intelligence analysts, enables them to make more consistent decisions. This approach has been embedded in a software tool and evaluated in a security intelligence analysis context. The developed tool relies on the Dominance-based Rough Set Approach (DRSA) (Greco et al, 2001). The DRSA is a preference learning approach extend classical rough set theory (Pawlak, 1982, 1991) to multicriteria classification. Rough set theory is a way of addressing analysis of imperfect data by taking lower (definitely belong) and upper approximations (possibly belong) of commonly held attributes between two objects. Figure 2 shows graphically a rough set M, its lower approximation M_* and its upper approximation M^*. The set difference $Bn=M^* \ M_*$ between M^* and M_* is called the boundary. The definition of approximation sets relies on the dominance principle (that can be seen as monotonicity constraints) stated as follows for our application: 'If report R_x is at least as risky as report R_y with respect to all relevant attributes, then report R_x should be classified at least as risky as report R_y'.

Figure 3 Principle of DRSA

3 This project is funded by the Centre for Research and Evidence on Security Threats (CREST, UK): ESRC Award number ES/N009614/1 received from the Economic and Social Research Council (ESRC, UK) and Centre for Research and Evidence on Security Threats (CREST, UK).

The DRSA has been broadly applied in a number of domains ranging from environment risk assessments, maintenance policy definition to an economic risk analysis. It is primarily used to assess data where decisions have been made and extract rules. Typically, this will be by taking an initial set of data with known outcomes or decisions (the learning set), applying an algorithm to extract rules and then applying these rules to predict the outcomes of new data. The advantage of this approach is that is means that the outcomes or decisions can be predicted for the new data set without

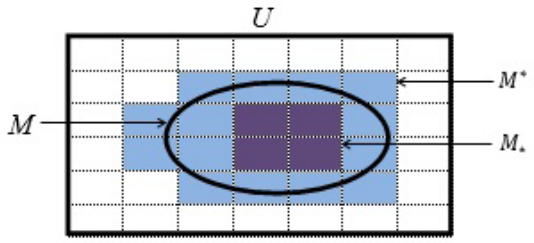

Figure 2
Lower and upper approximations of rough set M (Pawlak, 1991)

an extensive decision-making process. This principle is illustrated in Figure 3.

The input of DRSA is a decision table representing the description of a set of objects (reports) with respect to a set of condition and decision attributes. The entries of the decision table are attribute-value pairs. A generic representation of decision table is given in Table 1. The first column is the objects (intelligence reports in our application) number. The last column is the decision (risk level in our application) as expressed by the expert (analyst in our application).

In DRSA, the decisions are expressed on ordinal scale. In the considered application, we used a Likert scale of five levels from Cl_1=1 to Cl_5=5 where Cl_5 is the highest risk level. The other columns are condition attributes. The value for these attributes is extracted from the characteristics of the intelligence reports (such as location, frequency, etc.) or computed based on the scoring of intelligence reports by the analysts. More information about attributes extraction is given in Section 4.2.1.

Table 1 Representation of decision table

Object #	Attribute 1	...	Attribute n	Decision
R1	12	...	High	4
R2	5	...	Very high	3
R3	15	...	Moderate	4
R4	2	...	Low	1
R5	31	...	Low	5
R6	11		Moderate	2
R7	10		Moderate	2

In order to handle the monotonic dependency between conditions and decision at (risk level assignments), DRSA uses two collections of union of classes defined as follows:

- $Cl_t^{\geq} = \bigcup_{s \geq t} Cl_s$: upward union of classes;
- $Cl_t^{\leq} = \bigcup_{s \leq t} Cl_s$: downward union of classes.

where $Cl_t^{\geq}$ and $Cl_t^{\leq}$ are positive and negative dominance cones in decision space reduced to single dimension. Figure 4 shows graphically the definition of union of classes.

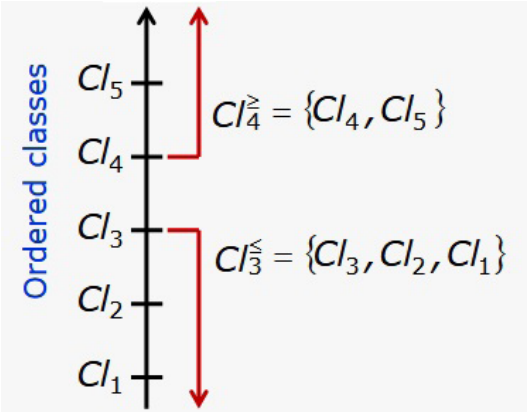

Figure 4 Definition of unions of classes

The main output of DRSA is a collection of decision rules. A decision rule is a consequence relation $E \rightarrow H$ (read as If E, then H) where E is a condition (evidence or premise) and H is a conclusion (decision, hypothesis). Each elementary condition is built upon a single condition attribute while a consequence is defined based on a decision attribute. Three types of decision rules may be considered in DRSA: (i) certain rules generated from lower approximations; (ii) possible rules generated from upper approximations; and (iii) approximate rules generated from boundary regions.

Only certain decision rules are considered here. Their general structures are as follows:

- **If** condition(s), **then** Risk Level $\geq Cl_t$
- **If** condition(s), **then** Risk Level $\leq Cl_t$

The condition part specifies values assumed by one or more condition attributes and the decision part specifies an assignment to one or more risk levels.

The quality approximations in DRSA are characterised by two basic measures. The first measure is the quality of approximation γ of partition **Cl**=$\{Cl_1,...Cl_5\}$ by means of condition attributes is defined as follows:
γ = *Correctly classified objects / Total number of objects in the system* (1)

The second measure is the accuracy α of the rough-set representation of classes computed as follows:

α=*Nb of objects in the lower approximation / Nb of objects in the upper approximation* (2)

When the upper and lower approximations are equal (i.e., boundary region is empty), then α=1, and the approximation is perfect. At the other extreme, whenever the lower approximation is empty, the accuracy is α=0.

The DRSA has been proven to be a particular efficient and effective machine learning approach, when compared to other methods such as regression analysis, other machine learning or data mining approaches. The main advantages are that it works with multiple data formats (e.g. text and numbers), can deal with gaps or inconsistencies in the data and is simple to understand using If-then rules. DRSA has been applied widely in non-defence domains and has been shown to provide significantly greater accuracy than other machine learning methods (e.g. nearest neighbours, support vector machine, decision trees).

However, the main advantages of DRSA in intelligence analysis context concerns its ability (i) to detect and dealt with inconsistent decision and (ii) to 'mimic' the analyst's behaviour. The ability of DRSA to detect and deal with inconsistent decision is due to the use of the dominance relation. This relation advocates that the risk level of a given intelligence report cannot be lower than the risk level of any other intelligence report representing similar or higher thread. The DRSA is also able to capture the preference and behaviour of analysts. In this respect, the DRSA is categorized by some authors as a preference learning method because it is used to build a preference model based on a sample of past decisions, via preference representation in terms of several If-then rules, for further prescriptive decision purposes. These rules should reproduce the behaviour of the analysts when applied to assess new and unseen intelligence reports.

3.3 INTELLIGENCE REPORTS ASSESSMENT

By embedding a DRSA capability within intelligence analysis software, it has the potential to significantly enhance how an intelligence analyst processes data and extracts intelligence. In a typical intelligence reports assessment exercise, the following basic steps apply (see Figure 5):

1. The analyst receives a series of reports from collection assets containing a number of attributes such as a unique identifier, time, date, source, geographic location, various text fields or imagery depending on the intelligence type. The analyst specifies terms of interest e.g. keywords, geographical areas, based on the objectives of the mission.
2. The data reports are presented in an ordered dataset, which includes hidden attributes that have been derived from the terms of interest, which the analyst specified. For example, this could be a simple count of analyst specified keywords within

each report. This also includes a decision-making attribute or dominance/risk level e.g. 1-5 priority or high/medium/low risk.

3. An initial risk level is given to each report based on the dominance relation. This will ensure that reports with similar attributes are within the same risk level group.
4. The analyst uses initial risk level (derived solely on the dominance relation) to prioritise a set of interesting reports to review. This represents the learning set. Once they have reviewed a report, they make a decision and allocate a risk level to the report. All reviewed reports that have been applied a risk level, represents the learning set.
5. The DRSA algorithm is automatically applied to generate classification If-Then rules, which are applied to the main data set. For example, if a report contains a combination of key words, occurs in a particular location and is of a certain source, it may be considered high risk.
6. The derived rules are automatically applied to the remaining reports that the analyst has not looked at yet, giving them a predicted DRSA derived risk level.
7. The analyst then orders or filters these non-viewed reports by risk level enabling them to be prioritised and for the high-risk reports to be brought to their attention. This speeds up the analyst's ability to search through data to identify critical information or intelligence, enhancing overall situational awareness. It also means they are less likely to waste time exploring intelligence reports that are not likely to be of interest to them.
8. The analyst repeats the previous steps until the mission is completed and they can provide an intelligence summary to their headquarters. The DRSA algorithm automatically update the risk levels, every time a new report is added and assessed or scored respectively. This ensures a dynamic updating of the rules and risk levels to reflect the analyst's latest decision-making process.

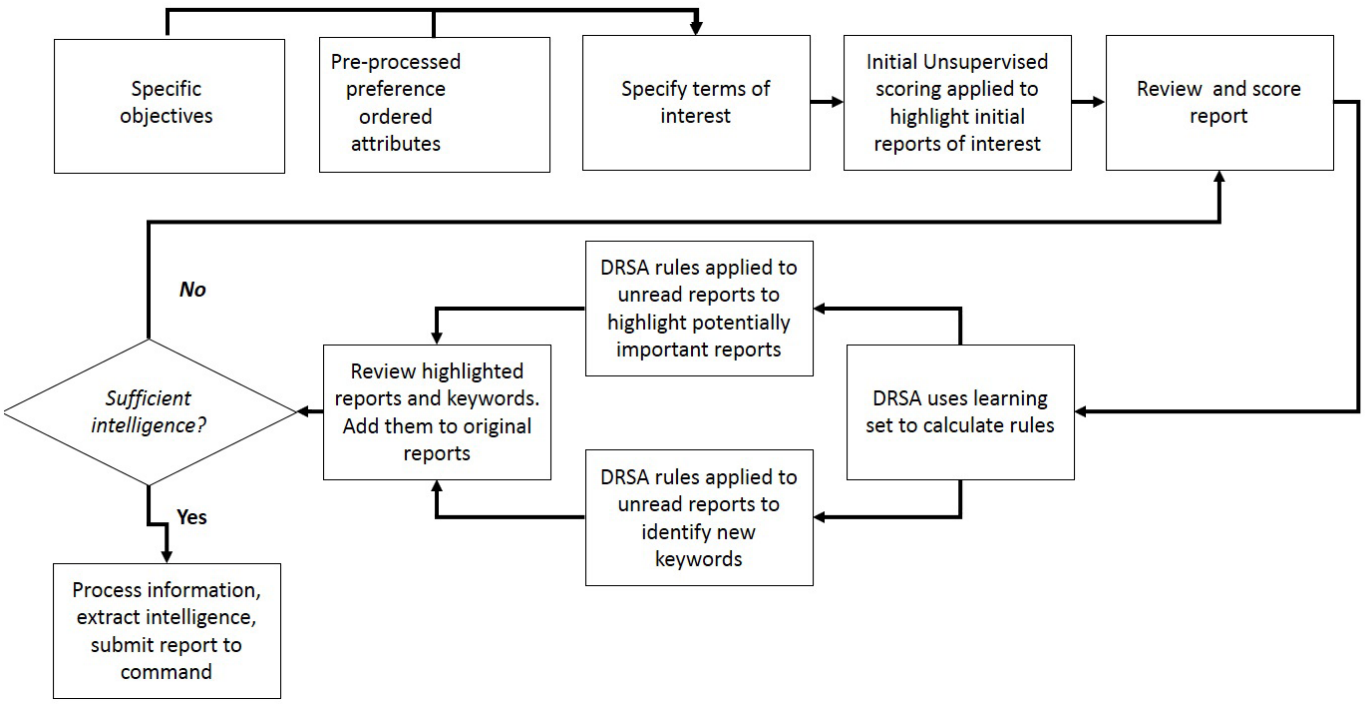

Figure 5 Flowchart of typical intelligence reports assessment exercise for single analysts

The main benefits of a DRSA enabled approach are:

1. Analysts can identify high priority or high-risk intelligence quickly without being distracted or wasting time reviewing low priority or low risk intelligence, improving the rate or automating the intelligence extraction process and enhancing situational awareness.
2. Support intelligence analyst training or when new analysts join an existing operation, as rules generated by an experienced analyst can be used and reviewed by more inexperienced analysts.
3. Using DRSA derived rules to improve/automate emitter signature detection-recognition-identification processes and communication network message sending priorities.

3.4 ACROSS-ANALYSTS AGGREGATION STRATEGY

Team based decision-making, or analysis of input from several analysts present a challenge to many machine/preference learning techniques due to the inherent inconsistency that multiple opinions can lead to. This is also the case in intelligence analysis are recognised by several authors as for example in Kerr et al (1996), Kerr & Tindale (2011) and Montibeller & von Winterfeldt (2015). Montibeller & von Winterfeldt (2015) advocate that some of the cognitive biases may be exacerbated at group level (or alleviated in some cases). The DRSA is well suited to dealing with such inconsistences at individual level but is not suitable for team decision-making contexts, such as those regularly operating in the intelligence community. Fortunately, recent advancements led to the design and development of the Dominance-based Rough Set Approach for Groups (DRSA*f*G) as an extension of DRSA to group decision-making context (Chakhar et al, 2016).

The DRSAfG algorithm permits to monitor individual decisions/reactions to data to create rules that are then used to score, sort and highlight existing and new data for all of the team. Scoring in the group context comes in two forms – individual and aggregated (or collective). Individual scores use just reports scored by the individual analysts, whereas collective scores merge the output of all analysts into a group view.

The basic idea for computing the aggregated score relies on the majority rule and veto effect (i.e., minority respect), which originated from Social Choice Theory and are now well established in decision-making (see, e.g., Roy, 1989; Bouyssou (2009)). The majority rule is based on decisions taken by a majority vote and veto effect is based on decisions taken by a minority vote. In Chakhar et al (2016), majority and veto rule are implemented trough the two following measures:

1. $C(R_i,Cl_j)$: concordance index representing the 'power' of analysts that agree on the assignment of intelligence report R_i to risk level Cl_j.
2. $D(R_i,Cl_j)$: discordance index representing the 'power' of analysts that do not agree on the assignment of intelligence report R_i to risk level Cl_j.

The computing of these measures uses a comprehensive weighting system that reflect the expertise and the quality of individual assignments given by each analyst as explained in the next subsection. Then, the assignment of intelligence report R_i to a risk level Cl_j holds if and only if:

$$\sigma(R_i,Cl_j)=C(R_i,Cl_j)*D(R_i,Cl_j)\geq, \quad (3)$$

where $\lambda \in [0.5,1]$ is the credibility threshold representing the minimum value for the credibility index $\sigma(R_i,Cl_j)$ for assigning an intelligence report R_i to risk level Cl_j. The assignment rule above ensures that intelligence report R_i is assigned to risk level L_j if and only if: (i) a majority of analysts, in view of their 'powers', support this assignment; and (ii) none of the analysts that do not support this assignment should express too strong disagreement. To identify the overall

risk level that should be assigned to a given report, we should first compute $\sigma(R_i,Cl_j)$ all risk levels. Then, an assignment interval $I(R_i)$ of the form $[l(R_i),u(R_i)]$, where $l(R_i)$ and $u(R_i)$ are respectively the lower and upper classes to which report R_i can be assigned, is deduced from the values of $\sigma(R_i,Cl_j)$. Finlay, if $l(R_i)=u(R_i)$], then overall risk level of intelligence report R_i is equal to $l(R_i)=u(R_i)$. Otherwise, some simple rules (such as minimum, maximum, median, floor and ceiling) are used to reduce the assignment interval $I(R_i)$ into a single risk level representing the final and overall score of the intelligence report.

3.5 INCORPORATING THE LEVEL OF EXPERTISE OF ANALYSTS

The DRSAfG has the facility to give different team members different levels of "power" which means their opinions hold more sway in the collective score. The definition of the levels of "power" of decision makers is a crucial step in group decision making as underlined by several authors (Chakhar & Saad, 2014; Cheng et al, 2018; Herowati et al, 2014, 2017; Zhang et al, 2014).

The authors in Chakhar & Saad (2014) enumerated several techniques to specify the weights in group decision making: (i) weights are defined explicitly by a mediator or an external independent person as in Leyva-Lopez and Fernandez (2003); (ii) weights are defined based on the hierarchical levels of involved decision makers; (iii) weights are defined explicitly using a given method as in Herowati et al (2014) and Yue (2011); and (iv) weights are inferred from input data using some form of regression as in Dias et al (2002).

Each of these techniques have some advantages and disadvantages and the selection of the technique to use is not obvious. The authors argue that the most important characteristic of weights definition method is the objectiveness of these weights. In this respect, it is advocated that the question is not how to use weights, but rather how to objectively quantify them (Cook, 2006).

Another important characteristic of weights definition methods is the ability of these methods to objectively measure the expertise of the decision makers or experts, as discussed in, e.g. Herowati et al (2014), Shanteau & Weiss (2014), and Weiss & Shanteau (2003). Indeed, and as pointed out by Herowati et al (2014), more experienced decision makers will generally provide more consistent decisions. This is confirmed in different real-world applications in which the authors were involved, as e.g., Mercat-Rommens et al (2015) and Saad & Chakhar (2009). With respect to intelligence analysis, Mandel & Barnes (2014) by studying the accuracy of forecasts in strategic intelligence, found that senior analyst not only did better than junior experts, they produced 68% of the forecasts despite constituting less than of the analysts.

An important aspect of the aggregation procedure introduced in the previous section is the use of comprehensive weighting system permitting to measure objectively the 'powers' of decision makers based on input data and scoring processes. In this weighting system, the contribution of each analyst to the 'collective' decision is measured by the quality of input data provided by this analyst. The weighting system combines both the quality of classification and the accuracy of the rough-set representation of risk levels. It enhances the weighting system used in Chakhar and Saad (2012), which is based on the quality of approximation only. In fact, the accuracy of the approximation of a single risk level, say Cl_t, obtained by a given analyst may be equal to 0 which means that analyst does not support any assignment of intelligence reports to risk level Cl_t. This fact is well supported by the proposed new weighting system. Using just the quality of approximation as weights does not allow to take into account this fact (since the quality of approximation characterizes the whole classification rather than the approximations of individual classes).

3.6 RESEARCH QUESTIONS AND HYPOTHESES

3.6.1 REDUCING THE TOTAL SCORING TIME AND IMPROVING THE EFFICIENCY AND EFFECTIVENESS

The assessment of intelligence reports relies on a preference learning method, namely DRSA. The working principle of DRSA is similar to classical machine learning methods: it uses a subset of data (reports in this application) to generate some rules permitting to generalise the preference information (which takes the form of scores assigned to reports here) to the whole dataset (all reports). This working principle will naturally reduce the computing time and improve the efficiency of analysts since they are not required to score all the intelligence reports to reach the same conclusion). These facts have been confirmed in the previous study reported in Baldwin et al (2016) where two groups of analysts reviewed the same set of reports, one group with DRSA and one without. Uses of DRSA were shown to gain a higher level of situational awareness more rapidly and spent more of their time reviewing 'relevant' data than the control group.

For the purpose of this paper, the reduction of the total scoring time and improvement of efficiency and effectiveness are evaluated through three measures, introduced hereafter. Let $n>0$ be the total number of reports received over a period T of time (e.g. duration of a session) and let T_R be the average processing time required to analyse an intelligence report by a given analyst. An analyst without DRSA will then require an average total processing time of $n \times T_R$ to score all the reports. By using DRSA, the analyst needs to score only a subset of $m<n$ reports in average; the remaining $n-m$ reports will be assessed automatically using the decision rules deduced form the scoring of the m reports. Then, the reduction of total scoring time can be measured as the difference between the allowed session time T and $m \times T_R$:

$$Total\ Scoring\ Time = T - (m \times T_R) \quad (4)$$

In practice, analysts are often under pressure and a non-DRSA analyst may not be able to assess all the reports during time period T, with potential security threat implications. This holds when the average total processing time exceeds the allowed processing time, i.e., $(n \times T_R)>T$.

The efficiency of an analyst is computed as the difference between 1 and the ratio of the number of reports (m) scored by this analyst during period time T by the maximum number of explicitly scored reports (M) during the same time period, i.e.:

$$Efficiency = 1 - (m / M) \quad (5)$$

The effectiveness analysts can be evaluated by examining patterns of 'hits', 'misses' and 'false alarms'. This can be achieved by analysing the final scores of intelligence reports with actual scored computed by the authors based on the scenario used during the evaluation exercise. In this application, correctly identifying intelligence reports with actual scores of 4 or 5 is considered as a hit while assigning a score of 1 or 2 to an intelligence report of actual score of 4 or 5 is considered as miss. False alarms holds when an

	Reports scored 4 or 5	*Reports scored 1 or 2*
Reports actually scored 4 or 5	Hits	Misses
Reports actually scored 1 or 2	False Alarms	Correct Rejections

Figure 6 Schematic representation of confusion matrix

intelligence report of actual score of 1 or 2 is scored 4 or 5. All these situations are summarized by the confusion matrix shown in Figure 6.

The situations shown in Figure 6 are formally designed in machine learning literature as follows:

- **True Positives** (i.e. Hits): The cases in which an analyst predicted 4 or 5 while the actual scores are 4 or 5.
- **True Negatives** (i.e. Correct Rejection): The cases in which an analyst predicted 1 or 2 while the actual scores are 1 or 2.
- **False Positives** (i.e. False Alarms or Type I Error): The cases in which an analyst predicted 4 or 5 while the actual scores are 1 or 2.
- **False Negative** (i.e. Misses or Type II Error): The cases in which an analyst predicted 1 or 2 while the actual scores are 4 or 5.

The accuracy for the confusion matrix can be calculated by taking average of the values lying across the main diagonal, i.e.:

Accuracy = True Positives + True Negatives / Total Number of Scored Reports (6)

In this present application, the accuracy for the confusion matrix will be used as a measure of effectiveness.

Our assumption is that the use of DRSA will permit to reduce the total scoring time and also improve the efficiency and effectiveness of decision. The first hypothesis to be tested is stated as follows:

Hypothesis H1: *The use of DRSA-based analysis will (i) reduce the average computing time of intelligence reports; and (ii) improve the efficiency and effectiveness of analysts.*

3.6.2 CORRECTION/REDUCTION OF NOISE AND BIAS ERRORS AT INDIVIDUAL LEVEL

Noise and bias errors at individual level arise mainly when intelligence analysts take decisions that deviate significantly from their own prior decisions. With respect to the current application, noise errors can be identified by analysing the scores successively provided by an analyst in different time points during the same work session or for the same project but in different time periods. Only the first case is considered in this paper. To measure the noise error, we will employ the well-known non-parametric statistics Kendall's tau (see Kendall, 1938; Nelsen, 2001), defined as follows. Let $(S_{1,t}, S_{1,t'})$, $(S_{2,t}, S_{2,t'})$,…, $(S_{n,t}, S_{n,t'})$ the scores of intelligence reports $R_1, R_2,...,R_n$ in time points t and t'. Then, any pair of scores $(S_{i,t}, S_{i,t'})$ and $(S_{j,t}, S_{j,t'})$ are said to be:

- concordant if (a) $S_{i,t} < S_{i,t'}$ and $S_{j,t} < S_{j,t'}$ or (b) $S_{i,t} > S_{i,t'}$ and $S_{j,t} > S_{j,t'}$.
- discordant if (a) $S_{i,t} > S_{i,t'}$ and $S_{j,t} < S_{j,t'}$ or (b) $S_{i,t} < S_{i,t'}$ and $S_{j,t} > S_{j,t'}$.
- neither concordant nor discordant otherwise.

Kendall tau $\tau_{t,t'}$ can be applied with or without ties (see e.g. Agresti (2010)). The second version of Kendall tau between the scores in time points t and t' is defined as follows:

$$\tau_{t,t'} = (n_c - n_d) / [(n_0 - n_1)(n_0 - n_2)]^{1/2} \quad (7)$$

where $n_0 = n(n-1)/2$; n_c is the number of concordant pairs; n_d is the number of discordant pairs; and $n_1 = \Sigma_k n_k (n_k-1)/2$; $n_2 = \Sigma_h n_h (n_h-1)/2$ with n_k and n_h are the number of tied values in the kth and hth groups of ties in the first and second series of scores, respectively. Note that there are other methods for measuring noise such as the one proposed by Kahneman, and Rosenfield (2016), where a noise index is computed as the difference divided by the average of any pair.

Kendall's tau lies in the range [-1,1]. If the agreement between the scores in two different time points is perfect (i.e. no noise error) it is 1. If the disagreement between the scores in two different time points is perfect, (i.e. scores are the reverse of each other and then all scores are noisy) it is -1. If the scores in two different time points were independent, then we would expect it to be approximately zero.

Similarly, bias errors can be identified by analysing the scores provided by an analyst in a given time point with the actual score of the report. If we denote by S_i^* $(i=1,...,n)$ the actual score of report R_i, then the bias errors between the scores in time point t and the actual scores can be measured using the Kendall tau between the pairs $(S_{1,t}, S_1^*)$, $(S_{2,t}, S_2^*)$,…, $(S_{n,t}, S_n^*)$. The description of the values of Kendall's tau in the range [-1,1] still applies here but it will concern the scores at a given time point and the actual scores.

Our assumption is that the use of DRSA will permit to reduce the number of noise and bias errors. The second hypothesis to be tested is then stated as follows:

Hypothesis H2: *The use of DRSA-based analysis will reduce the noise and bias errors at individual level.*

3.6.3 CORRECTION/REDUCTION OF NOISE AND BIAS ERRORS AT TEAM LEVEL THROUGH GROUP FEEDBACK

An important question that should be investigated in this study is the effect that group decision feedback support, as opposed to individual support, may have on the quality of decisions taken by an intelligence analyst team. In the considered application, the group feedback corresponds to the aggregated scores computed using the algorithm proposed by Chakhar et al (2016). In this paper, the effect of group feedback on the quality of decisions taken by an intelligence analyst team is evaluated using the Kendall's W (also known as Kendall's coefficient of concordance) non-parametric statistics (see e.g. Kendall & Babington Smith, 1939). Let S_{it} be the score of report R_i at time t where there are n reports and m (with $m>2$) time points. Then, the Kendall's W for noise errors is defined as follows):

$$W = 12S/m^2(n^3-n) \quad (8)$$

where: n is the total number of scored reports; m is the number of time points considered; and $S= \Sigma_{i=1,...,n} (S_i - \hat{S})^2$ with $S_i=\Sigma_{t=1,...,m} S_{it}$ is the total score given to report R_i, $\hat{S}=1/n\Sigma_{i=1,...,n,}$ and S_i is the mean value of these total scores. The Kendall's W for bias error is defined in the same way but it should also include the actual score of report R_i (which will give $m+1$ scores for each report).

If Kendall's W is 1, there is no noise or bias errors overtime. If Kendall's W is 0, then there is no overall trend of scores over time and scores may be regarded as essentially random. Intermediate values of Kendall's W indicate a greater or lesser degree of unanimity among the scores over time.

Our assumption is will reduce noise and bias errors at team level. Then, the following hypothesis will be tested during the validation phase.

Hypothesis H3: *The use of DRSA-based analysis with group feedback will reduce total scoring time as well as noise and bias errors.*

4. TOOL DESIGN AND DEVELOPMENT

4.1 TOOL DESIGN

Figure 7 provides a high-level illustration of developed decision tool. The intelligence reports (as .txt files) contain a number of attributes (time, date, frequency, transmitter, receiver etc.), which are then processed to extract additional attributes in order to present them in a structured data set. These additional attributes could include keyword frequencies, calculated distances from a key point, text length, but are also influenced by any search terms the user has specified. For example, this could include frequencies of key words or names, cover-terms and geographical areas of interest.

The structured data then has the DRSA algorithm applied to it. This creates an interest score, which drives the DRSA algorithm and leads to predicted interest levels for the remaining unseen reports. This enables the analyst to further prioritise how they search through these reports to extract intelligence. Throughout this process, the analyst constantly updates the search terms and scores new reports, which refreshes the interest scores derived from the DRSA algorithm.

4.2 SOFTWARE DEVELOPMENT

The software was developed in C++, which provided a flexible framework to explore and test different configurations for meeting the requirements and respond to user testing feedback. Three of the most important aspects that would influence the success of the tool were the attribute extraction, user interface and DRSA configuration.

4.2.1 ATTRIBUTE EXTRACTION

In order to enable the attribute extraction, the tool takes the original .txt file data and derives additional attributes by pre-processing the original report data to generate additional attributes configured to enable the DRSA algorithm. The original report attributes (e.g. time, source, frequency) could be used alone with the DRSA algorithm, however a greater number and fidelity of attributes is likely to improve the classification rules within the DRSA algorithm. This would lead to more accurate predicted interest levels

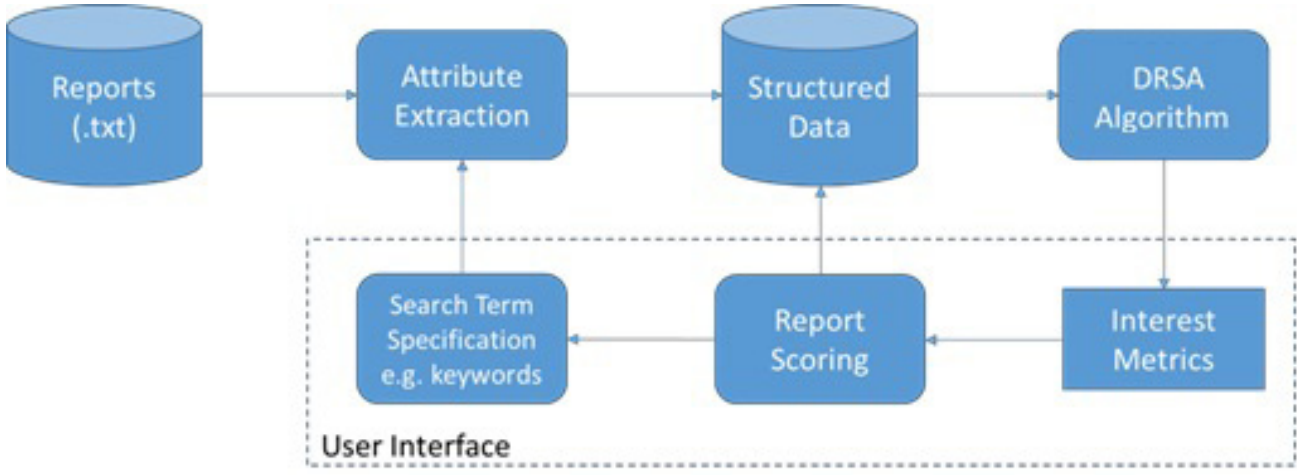

Figure 7 Tool design

for reports. To generate additional attributes, each .txt report would need pre-processing, in accordance with the search term specified by the analyst. These attributes would need to be in a dominance-based structure (i.e. there was a preference for a report with more or less of this attribute) to best apply the DRSA algorithm.

Various methods were explored deriving additional attributes to reflect search terms like keywords. This varied from a simple count attribute for each user-specified keyword to a single total keyword count attribute for all keywords. The former would likely enhance the report classification process and prediction accuracy but potentially be time-consuming to run, whilst the latter would likely lead to less accurate predictions but be more efficient in terms of processing time.

A compromise was reached, whereby three simple categories of keywords could be specified by the user (names, places, keywords) and preference ordered within a list. This would be used to derive three score attributes for names, places and keywords to enable more accurate report classification but not impacting significantly on processing time. This was further enhanced by using a weighting factor to allow analysts to prioritise keywords, with those word occurring at the top of the lists carrying more weight than those at the bottom. For example, a single mention of the most important person would carry more weight and likely lead to a greater name score attribute than multiple mentions of the least important person.

An example of the keyword specifications within the tool can be seen in left side of Figure 8.
A time-based attribute was also extracted and tested to reflect how an intelligence analyst's search terms of interest and priorities would change over time. Whilst in theory this would work, user testing showed that this time attribute became the dominant distinguishing attribute for all reports. This negated the ability of DRSA to accurately predict interest levels, so this attribute was removed. Following on

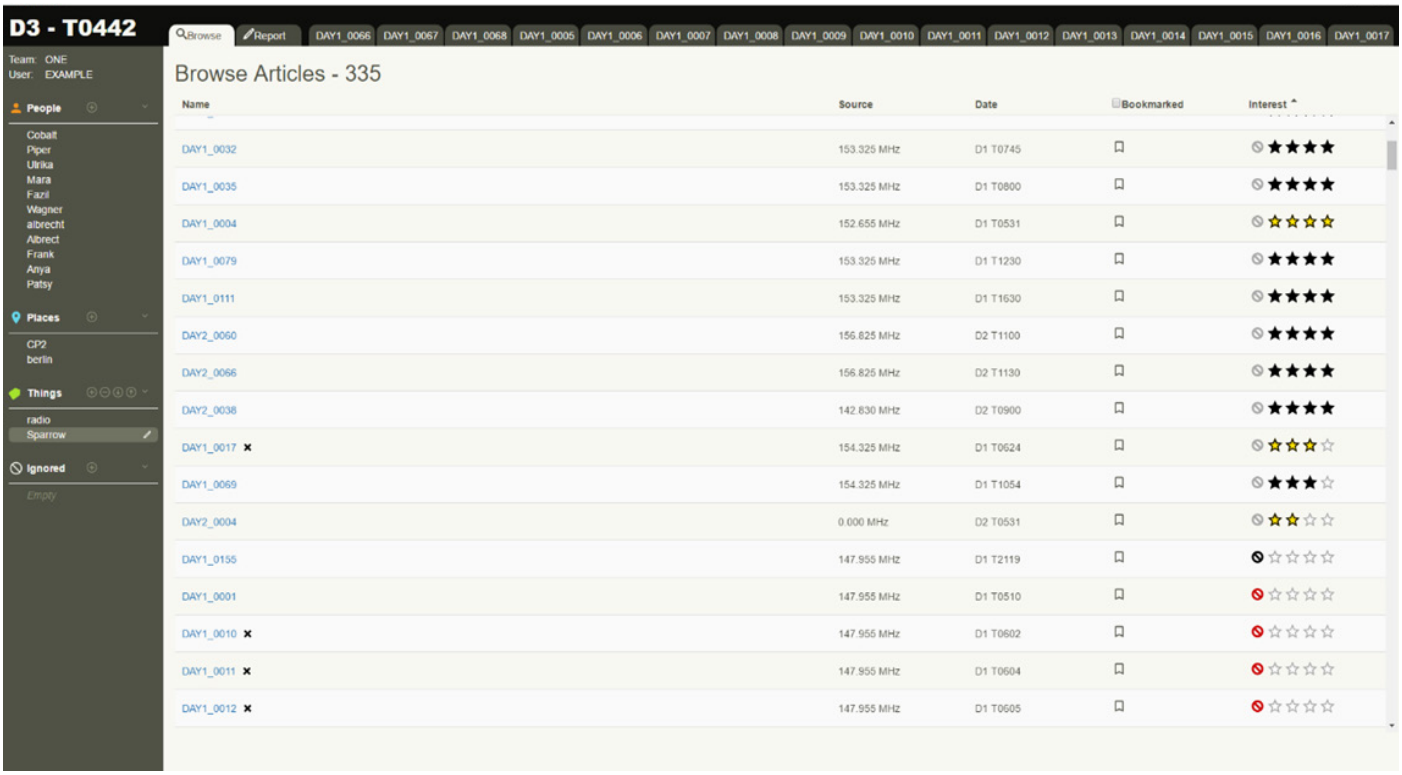

Figure 8 Keywords specification and scoring main interface

from the experiment, it was recognised that using time-bins or phases as an attribute would potentially work around this issue.

4.2.2 USER INTERFACE

The user interface was recognised as being important for ensuring the DRSA capability was implemented successfully. Key to this was ensuring users allocated an interest score to each of the reports they looked at to enable DRSA to function. This was done by enabling analysts to navigate their way through reports via a high-level summary table (see Figure 8). Analysts could double click on any report within the table to access the detailed attributes (including text content) and specify an interest score and new search terms (see Figure 9). The software ensured analysts could not close the report window without allocating a score.

4.2.3 DRSA CONFIGURATION

The way in which the DRSA algorithm were configured and interest scores indicated was also key to enabling an analyst to use the tool and take advantage of the predicted interest levels. Within the Report tab, an additional column was included reflecting the predicted DRSA interest levels. These were used by the analysts to identify and prioritise unseen reports that were potentially of high interest level.

It was recognised that DRSA can generate rules with multiple decision states i.e. an unscored report could potentially be of multiple interest levels. However, to minimise confusion for the analyst, the decision was taken to present a single DRSA derived interest level only; the highest predicted interest score. Whilst this could result in reports being predicted a higher interest level than they actually are (a false positive), it was felt that in an intelligence context this was better than selecting a lower predicted interest (a true negative) and potentially missing critical intelligence.

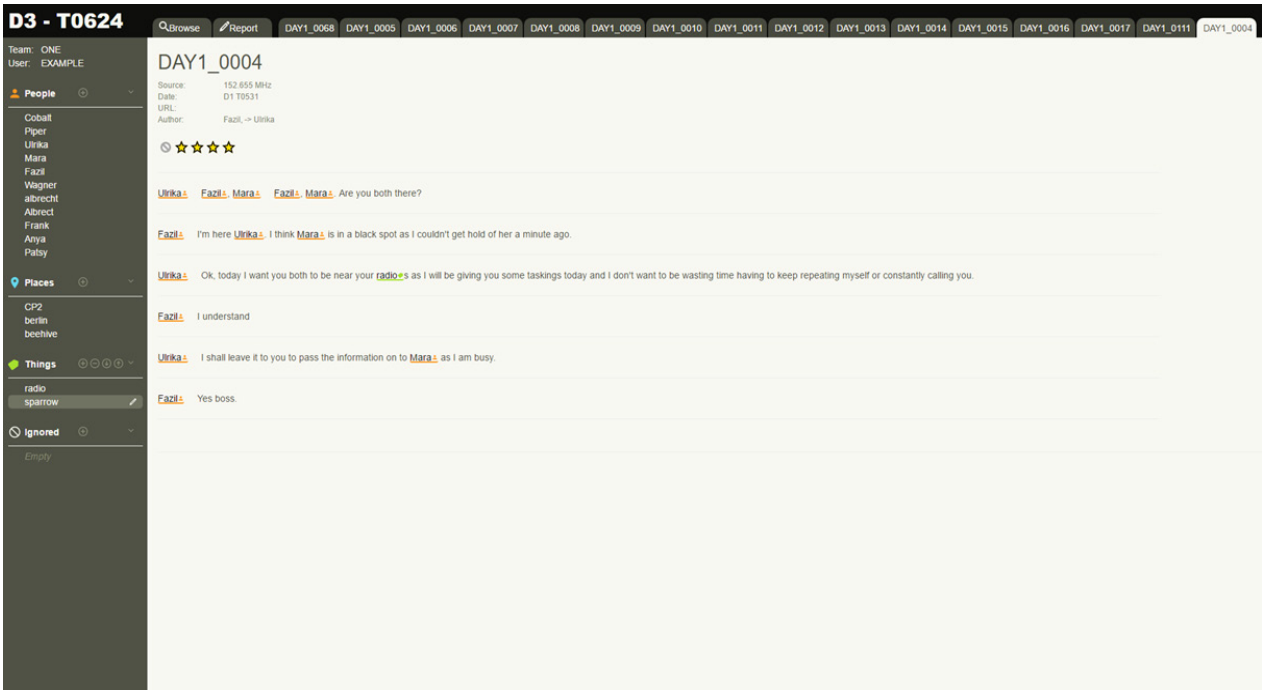

Figure 9 Example report window

5. TOOL VALIDATION AND ANALYSIS OF RESULTS

5.1 AIM AND DESCRIPTION

The aim of the developed tool is to help the security analyst to be: (a) effective i.e. to reach the right answer – a measure of effectiveness based on identification of critical target event information, measured by accurately scoring, and producing a summary report; and (b) efficient i.e. within the minimum time – an efficiency measure by which the decision is made (fewer reports assessed to reach the same conclusion). A series of trials involving simulated intelligence data were conducted to test the effectiveness and efficiency of the developed tool. The description of these trials is shown in Table 1. For the purposes of analysis, participants have been organized into three groups as follows:

(1) Non-DRSA: This is a control group where participants have no access to the DRSA. Participant of this group have the same main interface as the following two groups, but there is neither DRSA-based scoring nor the aggregated scores.
(2) DRSA without feedback: This group has access to DRSA-based scoring only. The aggregated score is hidden for the participants of this group;
(3) DRSA with feedback: This group has access to DRSA-based scoring as well the aggregated scores.

5.2 SCENARIO DESIGN

An example scenario was generated, which drew on the project team's experience of developing and supporting MOD SIGINT training exercises.

The scenario was designed to reflect a terrorist attack being planned in a Western European city and contained terrorist planning, logistics and reconnaissance cells. UK EW interception assets had captured radio transmissions of RED (terrorist), GREEN (military police) and WHITE (civilians) organisations over a five-day period and fed in their initial reports to a second line intelligence analysis cell.

These second line analysts were overwhelmed with this data and needed to process these reports to extract intelligence pertaining to the RED plans (see Figure 10).

Table 2 Characterises of conducted trials

Trial	Date	Group Number	Group Type	Number of Participants
1	29/03/2018	1	DRSA without feedback	7
2	04/06/2018	2	Non-DRSA, no feedback	8
2	04/06/2018	3	DRSA without feedback	7
2	04/06/2018	4	DRSA with feedback	7
3	14/06/2018	5	DRSA with feedback	8
4	31/07/2018	6	DRSA with feedback	8

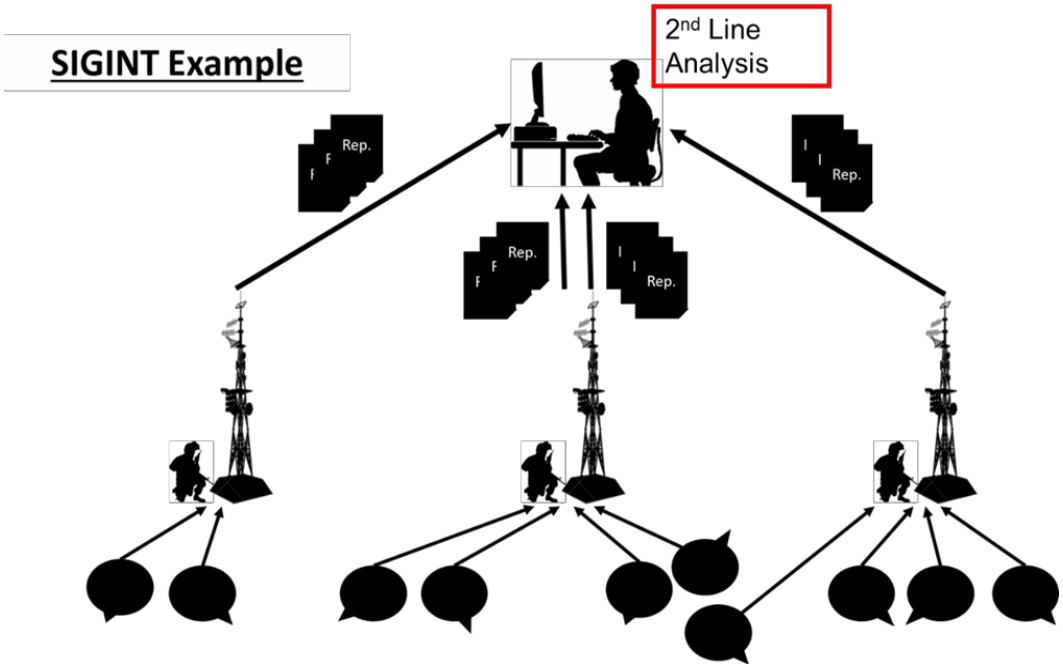

Figure 10 Scenario analyst scope

5.3 TEST DATA GENERATION

Over 450 example SIGINT reports were generated within an Excel spreadsheet and exported into individual .txt files to be used within the DRSA analysis tool. Each report contained a number of attributes as shown in Figure 11. In order to test the ability of DRSA to help the analysts quickly focus on the RED reports containing the intelligence, the majority of the reports generated were of GREEN and WHITE nature. The test data was modified to represent typical SIGINT report formats, which included removing some attributes (e.g. to/from, frequency) and including miss-spellings of names and places. Whilst this would make it more challenging to manually review the reports and extract intelligence, it was hoped this would help ability of DRSA to deal with data structure with missing attributes.

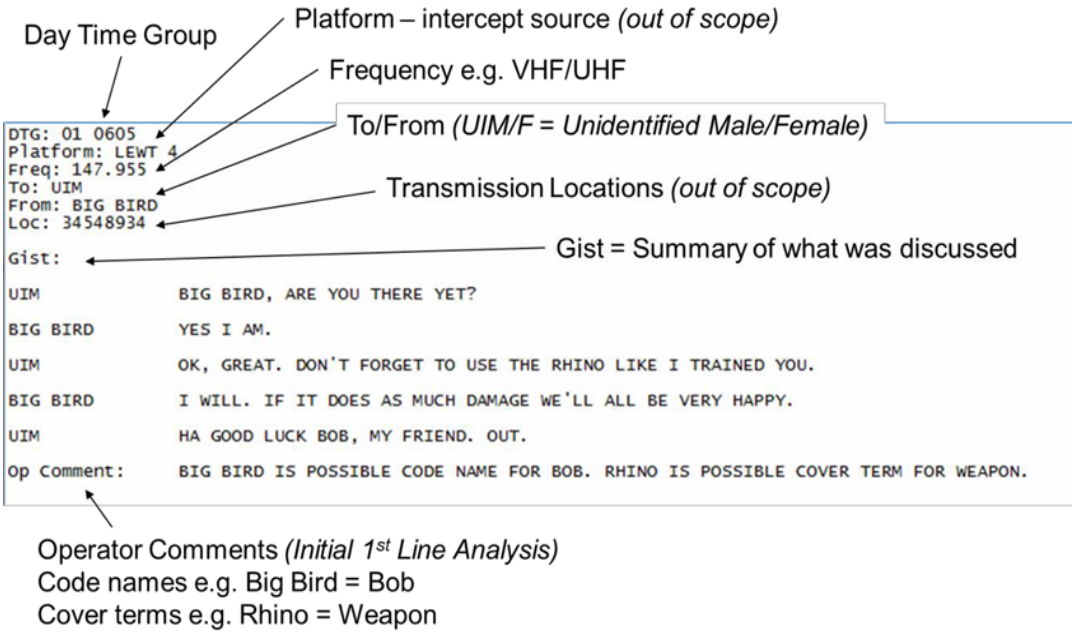

Figure 11 Example scenario report

5.4 RESULTS AND ANALYSIS

5.4.1 SOME GENERAL STATISTICS

We provide in this section some general statistics for all trial sessions. The considered statistics have been computed based on the results obtained at the end of the scoring process. These statistics concern (i) the number and percentage of received reports; (ii) number of explicitly scored reports; (iii) number of attributes used; (iv) quality of approximation; and (v) accuracy of assignments to risk levels. We mention that the quality of approximation is defined by Equation (1) as the ratio of all correctly scored reports to all reports and the accuracy of the rough-set representation of risk levels is computed by Equation (2) as the ratio between the number of reports in the lower approximation and the number of objects in the upper approximation. We note also that the quality of classification and accuracy for non-DRSA participants have been computed using all the scored reports as learning examples.

Table 3 provides the average values of above statistics. Based on the data in this table, we can conclude that: (i) participants using DRSA have scored much less reports (between 17.91% and 22.45%) than those with no DRSA (about 65%); (ii) none of non-DRSA participants have scored all the reports; (iii) non-DRSA participants have used slightly more attributes (about 98 attributes) than those using DRSA (between 59 and 84 attributes in average) about 98 versus ; (iv) there is no significant difference with respect to the number of criteria between participants using DRSA without feedback and those using DRSA with feedback; (v) the quality of classification and accuracy of DRSA participants are slightly better than those without DRSA; and (vi) there is no significant difference with respect to quality of classification and accuracy between participants using DRSA without feedback and those using DRSA with feedback.

5.4.2 REDUCING THE TOTAL SCORING TIME AND IMPROVING THE EFFICIENCY AND EFFECTIVENESS

5.4.2.1 Time Reduction Analysis

Equation (4) has been used to compute the reduction of total scoring time defined as the difference between the average total scoring time theoretically required to score n=441 reports and the average total scoring time of the reports actually scored by the analysts. The results of the non-DRSA group have been used to estimate the average processing time T_R (used in Equation (4)) of a single report as follows. First, for each member of the non-DRSA group we computed the average processing time of scoring a single report (T_R) as the ratio of the session duration (which is equal to 90 minutes for all participants) by the total number of scored reports. The average processing time required to score n=441

Trial	Group	Received reports		Scored reports		Number of criteria	Quality of approx.	Accuracy of assignments to risk level				
		Number	%	Number	%			1	2	3	4	5
1	DRSA without feedback	441	100	94	21.41	59	0.80	0.60	0.83	0.69	0.85	0.71
2	Non-DRSA	441	100	287	65.00	98	0.70	0.70	0.63	0.59	0.74	0.61
2	DRSA without feedback	441	100	89	20.12	84	0.80	0.73	0.68	0.83	0.63	0.88
2	DRSA with feedback	441	100	99	22.45	77	0.75	0.66	0.89	0.74	0.75	0.8
3	DRSA with feedback	441	100	79	17.91	82	0.92	0.85	0.92	0.83	0.88	0.73
4	DRSA with feedback	441	100	91	20.63	90	0.87	0.64	0.77	0.71	0.84	0.88

Table 3 Statistics about final results

reports is then obtained by multiplying the estimated value of T_R by the total number of reports n, i.e. n x T_R. The results of this exercise are summarised in Table 4, which shows an average processing time of 0.32 minutes per report and an average required total processing time of 140.87 minutes.

The reduction of total processing time is then computed through Equation (2). The obtained results are summarised in Table 5. This table shows an average processing time reduction by about 58.96 minutes.

5.4.2.2 Efficiency of Analysts

The efficiency of analysts as computed Equation (5) as the difference between 1 and the ratio of the number of reports (m) scored by these analysts during period time T by the maximum number of explicitly scored reports (M) during the same time period. The results are given in Table 5. We note that the maximum number of explicitly scored reports is M=287, which corresponds to average number of scored reports by non-DRSA participants.

We note that in the case of DRSA groups, the scores of all reports is atomically computed based on the rules extracted from the explicitly scored reports. Hence, the efficiency of analysts is computed with respect to non-DRSA participant. This means that values in the last column of Table 6 indicates the average efficiency ratio of analysts compared to non-DRSA participants. This also explains why the efficiency of non-DRSA participants is equal to zero in Table 6.

Participant #	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**Average actual scoring time of a single (T_R)**	**Average required total processing time (n x T_R)**
Session duration D (minutes)	90	90	90	90	90	90	90	90		
Number of scored report (n_s)	237	269	306	254	306	301	338	270		
Average scoring time per report (D/n_r)	0.38	0.33	0.29	0.35	0.29	0.30	0.27	0.33	0.32	140.87

Table 4 Estimation of average scoring time per report and a required total processing time

Trial	**Group**	Total number of reports (n)	Total number of scored reports (m)	Time Reduction (T-mT_R)
1	DRSA without feedback	441	94	59.92
2	DRSA without feedback	441	101	57.68
2	DRSA with feedback	441	89	61.52
3	DRSA with feedback	441	98	58.64
4	DRSA with feedback	441	103	57.04
		Average	97	58.96

Table 5 Calculation of processing time reduction

Trial	**Group**	**Number of explicitly scored reports (m)**	**Efficiency (1-m/M)**
1	DRSA without feedback	94	0.67
2	Non-DRSA	287	0
2	DRSA without feedback	89	0.69
2	DRSA with feedback	99	0.66
3	DRSA with feedback	79	0.72
4	DRSA with feedback	91	0.68
Average		123	0.57

Table 6 Efficiency of analysts at the end of the scoring exercise

5.4.2.3 ffectiveness of Analysts

The effectiveness of analysts is evaluated by examining patterns of 'hits' (i.e. True Positive), 'misses' (i.e. False Negatives) and 'false alarms' (i.e. False Positives) in the identification of important reports (i.e. reports scored either 4 or 5) by comparing the scores provided by the analysts with the correct scores. In this application, correct scores have been defined by the authors based on information provided by the senior intelligence expert that designed the scenario. The effectiveness of analysts is measured through the scoring accuracy (see Equation 6) and computed as the ratio of the number of True Positives plus the number of True Negatives by the total number of score reports. Effectiveness of analysts for the different session at the end of the scoring exercise is given in Table 7. The analysis of Table 7 indicates that the effectiveness of DRSA users are better than the effectiveness of non-DRSA users are more efficient. Table 7 also shows group feedback improves the effectiveness of analysts.

5.4.2.4 Discussion

Based on the results above, we can conclude that the use of DRSA permits the reduction of the scoring time and also improves the efficiency and effectiveness of decision. This confirms the Hypothesis H1 stated in Section 3.6.1: "*The use of DRSA-based analysis will (i) reduce the average computing time of intelligence reports; and (ii) improve the efficiency and effectiveness of analysts*". The reduction of scoring time is due to the machine learning aspect of the DRSA since analysts are asked to score a reduced set of intelligence reports, rather than the systemic scoring of all reports for non-DRSA users. An important remark is concerned with the minimum number of reports that should be scored by the analysts in order to enable DRSA to work properly. This is in fact not just specific to DRSA but also relates to all machine learning methods. With respect this question, the authors in Legay et al. (2015) identified some general guidelines that can be followed to obtain the 'best' set of learning examples: (i) the reports should be as representative as possible by including different specifications and characteristics; (ii) the reports should be non-redundant (in terms of their evaluation with respect to different attributes); (iii) the reports should cover all the risk levels; and (iv) the reports parts should ideally be well known to the decision maker/expert. The authors in Legay et al. (2015) also observe that there was no ideal theoretical number of learning examples. A limited number of learning examples might lead to a few and very generic decision rules and too great number of learning examples may lead to a high number of very specific and redundant decision rules.

The improvement of efficiency is also due the reduced number of reports that should be scored by the analyst. The rules inferred by DRSA from explicitly scored reports are automatically used to score all existing (or incoming reports). The improvement of effectiveness of analysts is due to the automatic application of decision rules. This will result to the elimination/reduction of the number of 'misses' (i.e. False Negatives) and 'false alarms' (i.e. False Positives). This is due to the 'algorithmic' behaviour of decision rules, which are applied with no need to human intervention. Indeed, several authors, e.g. Cheng et al, 2018; Kahneman &

Trial	Group	Total Number of Score Reports	Accuracy
1	DRSA without feedback	441	0.85
2	Non-DRSA	441	0.72
2	DRSA without feedback	441	0.83
2	DRSA with feedback	441	0.92
3	DRSA with feedback	441	0.88
4	DRSA with feedback	441	0.92

Table 7 Effectiveness of analysts at the end of the scoring exercise

Rosenfield, 2016), advocate that algorithms are more effective since they lead to the same decision if they are applied on the same or similar data while decisions specified by a human being may vary over time (e.g. as a result of time pressure, tiredness, loose of concentration, presence of external disturbances, work pressure, etc.). Although that Kahneman & Rosenfield (2016) support the idea that algorithms often lead to a reduction of noise and bias, they in the same time see the application of algorithms as a radical solution since they are politically or operationally infeasible. Here, we should mention that the use of decision rules is much more flexible than the use of classical formal algorithms since decision rules are inferred form the input of the analysts and also because the set of decision rules evolve over time, along the cognitive behaviour of the analysts.

5.4.3 IDENTIFICATION AND REDUCTION/CORRECTION OF NOISE AND BIAS ERRORS

5.4.3.1 Analyse of Noise and Bias Errors at Individual Level

Noise and bias errors are generally considered in team-oriented decision-making. It is also possible to identify noise and bias errors with respect to a single decision maker by considering the decisions she/he made over time. In the present case study, noise and bias errors at individual level are identified by analysing the scores successively provided by an analyst in different time points during the same scoring session. As indicated in section 3.6.2, noise and bias errors at individual level are measured using the non-parametric statistics Kendall's tau. Noise errors are then defined the agreement level between the scores given at two successive time points *t* and *t'*. A summary of noise error for different sessions and participants are given is given in Table 8. The analysis of Table 8 indicates that the use of DRSA reduces the noise errors. Table 8 also indicates that group feedback further reduces the noise errors.

Trial	Group	Kendall tau
1	DRSA without feedback	0.723
2	Non-DRSA	0.546
2	DRSA without feedback	0.683
2	DRSA with feedback	0.879
3	DRSA with feedback	0.915
4	DRSA with feedback	0.893

Table 8 Summary of noise errors

Bias errors are defined as the agreement level between the scores given at a given time point and the actual scores. As underlined earlier, correct scores have been defined by the authors based on information provided by the senior intelligence expert that designed the scenario. A summary of bias error is given in Table 9. The analysis of Table 9 indicates that the use of DRSA reduces the bias errors and indicates that group feedback further reduces the bias errors.

Trial	Group	Kendall tau
1	DRSA without feedback	0.656
2	Non-DRSA	0.437
2	DRSA without feedback	0.701
2	DRSA with feedback	0.82
3	DRSA with feedback	0.867
4	DRSA with feedback	0.812

Table 9 Summary of bias errors

The comparison of the figures in Table 8 and Table 9 shows that the DRSA is slightly more effective in reducing noise errors then bias errors.

5.4.3.2 Analysis of Noise and Bias Errors at Team Level

As indicated in section 3.6.3, noise and bias errors at team level are measured using the non-parametric statistics Kendall's W. A summary of noise error for different sessions and different groups are given is given in Table 10 while a summary of bias errors at group level is given in Table 11. The analysis of Table 10 and Table 11 indicates that group feedback reduces the noise errors. The comparison of the figures in Table 10 and Table 11 shows that the DRSA is slightly

more effective in reducing noise errors then bias errors, which confirms the same remark obtained with individual analysts.

Trial	Group	Kendall W
1	DRSA without feedback	0.805
2	Non-DRSA	0.722
2	DRSA without feedback	0.843
2	DRSA with feedback	0.858
3	DRSA with feedback	0.899
4	DRSA with feedback	0.807

Table 10 Summary of noise errors at team level

5.4.3.3 Discussion

Trial	Group	Kendall W
1	DRSA without feedback	0.688
2	Non-DRSA	0.795
2	DRSA without feedback	0.7
2	DRSA with feedback	0.735
3	DRSA with feedback	0.923
4	DRSA with feedback	0.784

Table 11 Summary of bias errors at team level

Based on the results above, we can conclude that the use of DRSA permits the reduction of noise and bias errors, confirming thus the Hypothesis H2 stated in Section 3.6.2: "*The use of DRSA-based analysis will reduce the noise and bias errors at individual level*". The use of DRSA significantly reduces noise and bias errors because a basic assumption in DRSA is that if we have two reports R_1 and R_2 such that the evaluations of R_1 on the all attributes are equal or worst then those of report R_2, then report R_1 should be assigned a higher risk level than report R_2. Based on this assumptions, the DRSA can identify two types of noise error that occur (i) when two reports with same description (i.e. with same values on all attributes) are assigned to two different risk levels; or (ii) when two reports with different description (i.e. one of them is worst or best on all attributes then the other) are assigned to the same risk level. If during the scoring process the analysts fail to respect these rules, then the quality of approximation will necessarily less than one, indicating the presence of inconsistency. The analysts can then revise her/his assignments to correct these noise errors. We note that it is easy to correct this type of inconsistency within the developed tool since quality of approximation is computed after any modification of in scores of any report, so analyst can identify the cause of the inconsistency on-the-fly. If the analyst did not revise her/his assignment, the DRSA can then reduce this error since the basic DRSA (which implemented in the current tool) identifies the different risk levels of similar reports specified by the analysts and then pick the highest one for all of them. We note that a more recent version DRSA proposed in Blaszczynski et al (2007) and instead of selecting the highest risk level provided by the analyst and assign all of them, uses more advanced rules to identify the risk level of a given report. However, the solution used the basic DRSA is more prudent and more appropriate in the considered application.

The identification and measurement of bias errors require the availability of the correct answers (scores in our case) but, as remarked by Kahneman & Rosenfield (2016), correct answers will be known at the end of mission, if ever. This was possible in the conducted exercise since the correct scores were computed by the authors based on input from the scenario designer. In practice, however, it is difficult to know in advance the correct answer or in real-time during the scoring process; correct answers will normally be discovered at the end of the mission. Due to this fact, techniques and strategy for bias errors reduction should be oriented to the reduction or elimination of the sources that may lead to bias errors instead of trying to handle bias errors themselves. The developed tool can be enhanced better anticipate the correct answers in two different ways. First, by enriching the learning set by some relevant historical reports (that should concern similar missions then the one under investigation). These historical reports can serve as benchmarks and will be used by the DRSA to identify the relevant reports faster based

on the rules deduced from these historical reports in the beginning of the scoring process. Second, historical reports can be used as testing sets to evaluate the quality of current scoring process.

The results above also confirm Hypothesis H3 stated in Section 3.6.3: "*The use of DRSA-based analysis with group feedback will reduce total scoring time as well as noise and bias errors*". The main explication of noise and bias errors reduction at team level the 'framing' effect since it has been shown (see e.g. Zhang et al, 2014) that individuals in a team-oriented analysis tend to follow the group (aggregated) decision, which will speed up the processing time. In proposed tool, the noise and bias errors reduction relies on the majority rule implemented by the aggregation procedure. Indeed, the aggregation procedure tends to bring together the most frequent decisions and so reduce their dispersion.

6 CONCLUSION

The proposed research contributes to the current theoretical knowledge base with respect to decision support analysis for improving quality of judgements. More specifically, the empirical evaluation of judgements will contribute to our understanding of bias and noise errors inherit when making judgments. To the best of our knowledge, the application of the concepts of concordance and discordance (which originated from Social Choice theory and are now well established multicriteria analysis; see Chakhar et al, 2016) to implement the majority principle and veto effect in a group decision making context has never been used in a security intelligence framework. Furthermore, we will be able to propose, and validate, more advanced capabilities to better exploit the outputs of the DRSA.

The output from this work is a validated decision support tool that can be used within the intelligence analysis community to better understand the consistency of their judgements and enable them to extract more useful intelligence more efficiently. This will improve the ability to detect potential threats and mitigate them with effective deployment of resources, reducing the risks and overall security threat from criminals, terrorists and military opponents. In recent years there has been a huge drive on investing in new sensors (e.g. CCTV, UAVs) and collecting open-source data, which has resulted in significant volumes of data that is challenging to process and extract useful intelligence from in a timely manner. This is evidence now by the variety of calls, most commonly within the Defence Community2, to address this issue.

Specifically, the DRSA tool could be used to:

1) Inform training practices – by capturing what behaviours drive different judgements between analysts e.g. What attributes (factors) do "good" analysts use to make their decisions?

2) Improve decision making – by being embedded as a decision aide alongside existing intelligence analysis tools to provide feedback on the consistency of the analysts' judgements.

3) Inform new standards for intelligence processing - for example, informing the new standards to be drawn up by the College of Policing.

ACKNOWLEDGEMENTS

This project considered in paper has been funded by the Centre for Research and Evidence on Security Threats (CREST, UK): ESRC Award number ES/N009614/1 received from the Economic and Social Research Council (ESRC, UK) and Centre for Research and Evidence on Security Threats (CREST, UK). The authors are then very thankful to ESRC and CREST for supporting this study. The authors would also like thank all participants for their valuable inputs to this project.

7 REFERENCES

1. Adame BJ. Training in the mitigation of anchoring bias: A test of the consider-the opposite strategy. Learning and Motivation, 53, 36-48, (2016).
2. Agresti, A. (2010). Analysis of ordinal categorical data (Second ed.). New York: John Wiley & Sons.
3. Arkes, H.R., Blumer, C. (1985) The psychology of sunk cost. Organizational Behavior and Human Decision Processes 35(1):124–140.
4. Baldwin, T., Shimell, J., Andrews, S., Davies, E., Malinowski, M., Marklew, T., Oulhadj, D., Chakhar, S., CDE (Centre for Defence Enterprise), Applying a Dominant Rough Set Approach to improve intelligence data processing and enhance situational awareness, CDE42166, Reference J485-DRSA-01, 29/06/2016.
5. Blaszczynski, J., Greco, S., Slowinski, R. (2007) Multi-criteria classification - A new scheme for application of dominance-based decision rules, European Journal of Operational Research, 181(3):1030-1044.
6. Bouyssou, D. (2009) Outranking Methods. In: Encyclopedia of Optimization, Floudas, C.A, Pardalos P.M., editors, Springer US, pp. 2887-2893.
7. Chakhar, S., Ishizaka, A., Labib, A, Saad, I., Dominance-based Rough Set Approach for Group Decisions, European Journal of Operational Research, 251:206-224, (2016).
8. Chakhar, S., Saad, I. (2012) Dominance-based rough set approach for groups in multicriteria classification. Decision Support Systems 54(1):372-380.
9. Chakhar, S, Saad, I. (2014). Incorporating stakeholders' knowledge in group decision-making, Journal of Decision Systems, 23(1):113-126.
10. Cook, W.D. (2006) Distance-based and ad hoc consensus model in ordinal preference ranking with intensity of preference, European Journal of Operational Research 172(2):369-385
11. Cheng, D., Zhou, Z., Cheng, F., Wang, J. (2018) Deriving heterogeneous experts weights from incomplete linguistic preference relations based on uninorm consistency. Knowledge-Based Systems 150:150-165.
12. Cox, J, Nguyen, T, Thorpe, A, Ishizaka, A, Chakhar, S., Meech, L 2016, 'A decision rule approach for analysing the attractiveness of crowdfunding projects'. OR58 Annual Conference, Portsmouth, United Kingdom, 6/09/16 - 8/09/16, pp. 1-20.
13. DeRosa, M., Data mining and data analysis for counterterrorism, Center for Strategic and International Studies (CSIS),Washington, March, 2004.
14. Dias, L.C., Mousseau, V., Figueira, J.R., Climaco, J. (2002) An aggregation/disaggregation approach to obtain robust conclusions with ELECTRE TRI. European Journal of Operational Research 138:332-348.
15. Fischhoff, B., Chauvin, C., editors. Intelligence analysis: Behavioral and social scientific foundations. Washington, DC: National Academies Press; 2011.
16. Greco, S., Matarazzo, B., Slowinski, Rough sets theory for multicriteria decision analysis. European Journal of Operational Research,129(1),1–47, (2001).
17. Hammond, J. S., Keeney, R. L., Raiffa, H. The hidden traps in decision making. Harvard Business Review, 84, 118–126, (2006).
18. Herowati, E., Udisubakti, C., Joniarto Parung S. (2017). Expertise-based ranking of experts: An assessment level approach, Fuzzy Sets and Systems, 315: 44-56.

19. Heuer Jr., R.J. Psychology of intelligence analysis. Washington, DC: Central Intelligence Agency, Center for the Study of Intelligence; 1999.

20. Heuer, R.J., Randolph H.P. (2015) Structured Analytic Techniques for Intelligence Analysis, 2nd Edition, Sage.

21. Heuer Jr, R.J., Heuer, R. J., Pherson, R. H. (2010). Structured analytic techniques for intelligence analysis. Cq Press.

22. Hills, M., Sociotechnical gambits that destroy cyber security and organisational resilience. In:

23. Hills, M. (ed.) Why Cyber Security is a SocioTechnical Challenge: New Concepts and Practical Measures to Enhance Detection, Prevention and Response. New York: Nova Science Publishers, (2016).

24. Hope, L., Blocksidge, D., Gabbert, F., Sauer, J. D., Lewinski, W.,Mirashi, A., Atuk, E. Memory and the operational witness: Police officer recall of firearms encounters as a function of active response role. Law and Human Behavior, 40, 23–35, (2016

25. Hu, Q., Chakhar, S., Siraj, S., Labib, A., Spare parts classification in industrial manufacturing using the dominance-based rough set approach, European Journal of Operational Research. 262(3):1136-1163

26. Kahneman D. Thinking, Fast and Slow, Farrar, Straus and Giroux, New York, (2011).

27. Kahneman D., Knetsch, J.L., Thaler, R.H. Anomalies: The endowment effect, loss aversion, and status quo bias. Journal of Economic Perspectives, 1991; 5(1):193–206

28. Kahneman, D., Rosenfield, A.M., Noise: How to overcome the high, hidden cost on inconsistent decision making, Harvard Business Review, 94/10, 38-46, 2016.

29. Kaplan, E. Operations Research and Intelligence Analysis, in Fischhoff, B.; Chauvin, C., editors. Intelligence analysis: Behavioral and social scientific foundations. Washington, DC: National Academies Press; 2011

30. Kendall, M. (1938). A New Measure of Rank Correlation. Biometrika 30(1-2):81-89.

31. Kendall, M. G., Babington Smith, B. (1939). The Problem of m Rankings. The Annals of Mathematical Statistics, 10 (3): 275–287

32. Kerr, N.L., MacCoun, R.J., and Kramer, G.P. (1996) Bias in judgment: Comparing individuals and groups. Psychological Review 103(4):687–719.

33. Kerr, N.L. and Tindale, R.S. (2011) Group-based forecasting?: A social psychological analysis. International Journal of Forecasting 27(1):14-40.

34. Legay,. C., Cloutier, G. , Chakhar, S. , Joerin, F., Rodriguez, M.J. (2015) Estimation of urban water supply issues at the local scale: a participatory approach. Climatic Change, 130(4):491-503

35. Leyva-Lopez, J.C. Fern\'andez-Gonz\'alez, E. (2003). A new method for group decision support based on ELECTRE III methodology. European Journal of Operational Research, 148(1):14-27.

36. Mandel, D.R. and Barnes, A. (2014). Accuracy of forecasts in strategic intelligence.

37. Proceedings of the National Academy of Sciences. National Academy of Sciences, 111(30):10984-10989.

38. Mercat-Rommens, C., Chakhar, S., Chojnacki, E., Mousseau, V. (2015). Coupling GIS and multi-criteria modelling to support post-accident nuclear risk evaluation. In: Evaluation and decision models with multiple criteria: case studies. Bisdorff, R., Dias, L. C., Mousseau, V., Pirlot, M. & Meyer, P. (eds.). Berlin: Springer, pp. 401-428 (International handbooks on information systems series).

39. Montibeller, G., Von Winterfeldt, D. (2015). Cognitive and motivational biases in decision and risk analysis. Risk Analysis, 35(7), 1230-1251.

40. Nelsen, R.B. (2001), Kendall tau metric, in Hazewinkel, Michiel, Encyclopedia of Mathematics, Springer Science+Business Media B.V. / Kluwer Academic Publishers.

41. Pawlak, Z. (1982) Rough sets. International Journal of Information & Computer Sciences, 11:341-356.

42. Pawlak, Z. (1991) Rough set. Theoretical aspects of reasoning about data. Dordrecht: Kluwer Academic Publishers.

43. Reyna V.F., Chick, C.F., Corbin, J.C., Hsia, A.N. (2014). Developmental Reversals in Risky Decision Making: Intelligence Agents Show Larger Decision Biases Than College Students, Psychical Science, 25(1): 76–84.

44. Roy, B., Main sources of inaccurate determination, uncertainty and imprecision in decision models. Mathematical and Computer Modelling 12 (10/11), 1245-1254, (1989).

45. Saad, I., Chakhar, S. (2009). A decision support for identifying crucial knowledge requiring capitalizing operation. European Journal of Operational Research, 195(3):889-904.

46. Shanteau, J., Weiss, D.J. (2014). Individual expertise versus domain expertise. American Psychologist, 69(7):711-712.

47. Shanteau, J. (1992). Competence in experts: The role of task characteristics. Organizational Behavior and Human Decision Processes, 53:252-266.

48. Sigurdsson, E.H., Understanding faulty managerial decision-making: Why do company executives make poor strategic decisions? a multiple-case study approach, Thesis, Aarhus University, Denmark, (2016).

49. Tversky, A., Kahneman, D., Judgment under uncertainty: Heuristics and biases, Science, 185, 1124-1131, (1974).

50. Tversky A, Kahneman D. The framing of decisions and the psychology of choice. Science, 1981; 211(4481):453–458.

51. Weiss D.J., Shanteau J. (2003). Empirical assessment of expertise. Human Factors, 45(1):104-116.

52. Yue, Z. (2011). A method for group decision-making based on determining weights of decision makers using TOPSIS. Applied Mathematical Modelling, 35(4):1926-1936.

53. Zhang, B., Dong, Y., Xu, Y. (2014). Multiple attribute consensus rules with minimum adjustments to support consensus reaching. Knowledge-Based Systems, 67:35-48.

PART III: SOFTWARE AND EXPERIMENT PLAN DETAILS:

EXPERIMENTS DESIGN & SYSTEM SPECIFICATION

1. AIM OF THE EXPERIMENTS AND SYSTEM:

The Dominance-based Rough Set Approach (DRSA) is a well-known multicriteria classification method (see Appendix A). We have successfully used this method to design, implement and test a tool[4] in order to help analysts to process large volumes of data by capturing the rules and attributes by which they prioritised the data (intelligence reports). The previously developed tool has been designed for single analysts. The present project extends the existing tool by adding group decision-making capabilities and by incorporating group feedback. In this report, we design a series of experiments to test whether the extended DRSA Team tool might improve individual performance. More specifically, the aim of the developed tool is to help the security analyst to be:

a) Effective i.e. to reach the right answer – a measure of effectiveness based on identification of critical target event information, measured by accurately scoring, and producing a summary report.

b) Efficient i.e. within the minimum time – an efficiency measure by which the decision is made (fewer reports assessed to reach the same conclusion).

BRIEF ABOUT THE DESIGN OF THE EXPERIMENTS:

The current research will assess the performance of individual analysts (IA) with respect to effectiveness and efficiency where the task is to identify and prioritise intelligence information pertaining to a critical target event (Experiment 1). In Experiment 2, we will examine the extent to which effectiveness and efficiency of individual decision-making is impacted by an awareness of group level decision-making for the same target material.

DATA SETS USED:

- The organisers (Polaris) has a set of reports, where it is already known which ones are critical and which are not. This assessment information is not known t participants.
- We will assume that if participants rank a report as 4 or 5 (on a scale of 1-5) then the information is critical. Otherwise, any other ranking is considered non-critical.
- The organisers also know the right answer for the summary report.

All the above information will help us to measure effectiveness.

Briefly the scenario of the experiment concerns a critical target event (in this instance, a terrorist plot) which will not be revealed to the participants other than through a large set of intelligence reports. Participants will assess these intelligence reports which include information that is both relevant and irrelevant with respect to the target event. The goal for participants is to

4 In a project funded under the UK Ministry of Defence (MoD) Centre for Defence Enterprise (CDE) Autonomy and Big Data competition.

extract relevant intelligence knowledge from the data by judging the degree of interest/relevance of each report. Having completed the judgment task, all participants will submit a final summarised report addressing key questions (What, Who, Where, How, and When) with respect to the critical target event. These answers should characterise the nature of the plot (for the purpose of testing the methodology, it is assumed here that all questions are equally weighted). The performance of each IA, or the group (N=30 for Experiment 1 and 40 for Experiment 2), will be measured by the accuracy of responses to the key questions and the number of reports analysed. This measure can be captured by dividing number of critical reports judged by the time elapsed between the beginning of the session and the sending of the final reports with the answers.

Therefore, we will be able to capture two measures:

1. The proportion of critical target reports each participant has worked on by the end of the experiment (i.e. how many reports relating to the critical incident they have processed.
2. The number of reports in total processed by an individual within a particular time frame is a feature of their individual speed of work (i.e. how many they work through). Here, please note that this measure is different in meaning to the number of critical reports they've accessed. We hypothesis that efficiency equals a percentage expressed as ' A divided by B'.

This measure will also be applied to see how each participant is efficient with respect to processing critical reports in a given time frame.

In terms of Dependent Variables (DV) for both experiments, the main outcome measure in the study is the extent (in terms of effectiveness and efficiency) to which the analyst identifies key features of the plot. This will be quantified through the following performance measures: 1) how many answers to the summary report in terms of What/who/where/how/when questions were in fact answered correctly at the end of the simulation (an effectiveness measure), and 2) the number of reports processed in order to identify the target information (an efficiency of the decision making process measure).

The scoring given by each participant will be used to compute the degree of noise.; where noise is measured as the degree of scatter of the judgements using the noise-index indicator proposed by Kahneman and Rosenfield (2016)[5], as well as other statistical measures as explained in Appendix 1.

5 Noise Index: A noise index for each case, which answered the following question: "By how much do the judgments of two randomly chosen Decision Makers (DM) differ?" This amount is expressed as a percentage of their average. Suppose the assessments of a case by two DMs are £600 and £1,000. The average of their assessments is £800, and the difference between them is £400, so the noise index is 50% for this pair. We performed the same computation for all pairs of DMs and then calculated an overall average noise index for each case. In our case we use ordinal scale (from 1 to 5) instead. So, we will need to be map the ordinal to cardinal scale.

2. MORE DETAILS ABOUT SYSTEM ARCHITECTURE AND DATA ANALYSIS:

This section provides more details about both the system architecture and how results will be analysed.

BACKGROUND ABOUT SOFTWARE DESIGN AND SYSTEM ARCHITECTURE:

- Each analyst will have his/her own session as a client to a central system.
- There is no need that the analysts work together. This means that Mark [Polaris] just needs to add (i) the aggregation procedure (Note: This procedure extends the well-known multicriteria classification method DRSA to supports multiple decision makers. Brief presentations of DRSA and the aggregation procedure are given in Appendices 2 and 3, respectively); (ii) a way to collect/send scoring between individual analysis and a central server; and (iii) add a Summary Tab to the main screen permitting Individual analysts specify and send their final reports to the central server. The architecture of the system is as in Figure 1. Note that the aggregation procedure requires inputs from at least two different individual analysts. More details are provided in the Appendix 3.
- The group decision making aspects will be gathered centrally in an indirect way through the aggregation procedure, which will be located in the central server.
- Individual analysts are not required to have any prior knowledge about the aggregation procedure, but they will be informed during experiment 2 that "the group aggregation procedure will allow their individual scores of their rankings to be combined".
- The software should be parameterised in the sense that it can permit to display an additional column

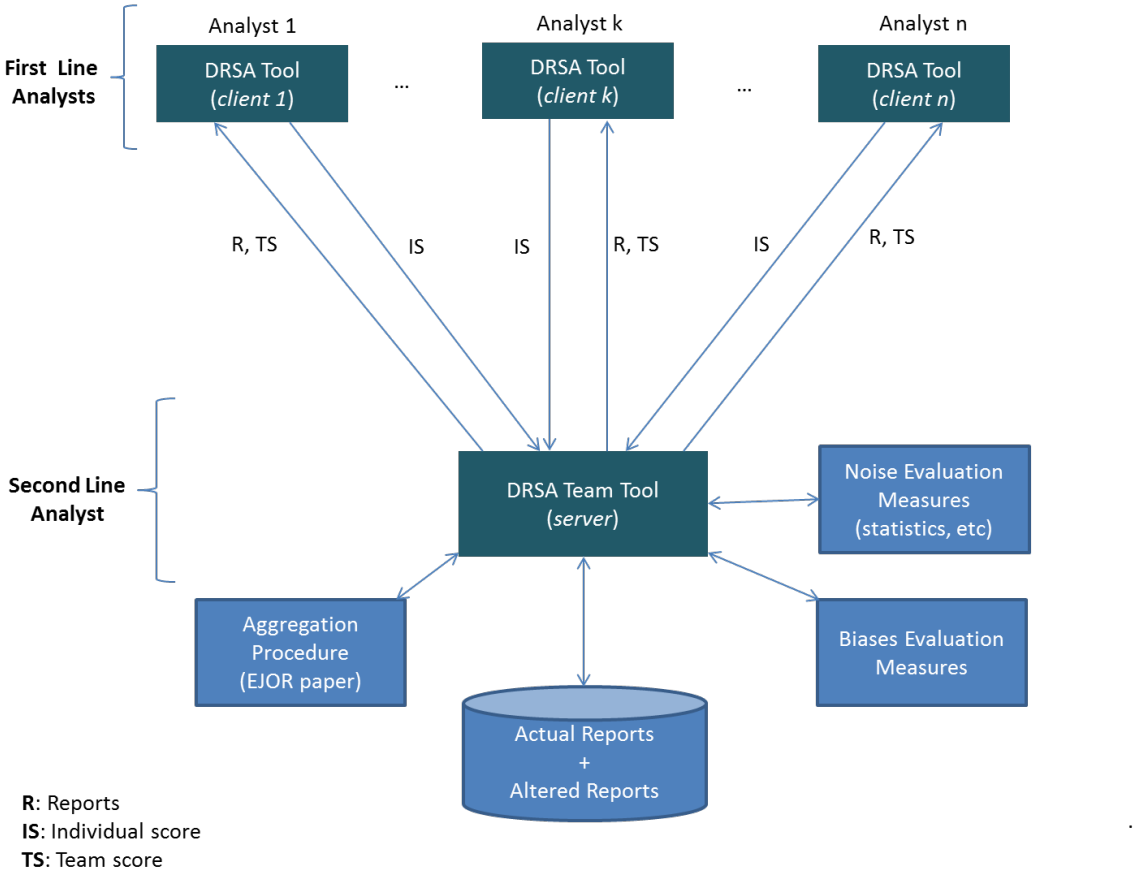

Figure 1: System Architecture T

in the scoring screen of each Individual analysts about group score, or to hide that column.

- The data will be collected over regular intervals of time (or, alternatively, after a given number of score reports).
- Will we know how many of the critical (as opposed to lure) reports each participant has analysed by the end of the exercise.
- All reports fully randomised in terms of presentation.
- At the end of the session, each analyst will provide the software with a final summary. The final summary should be a brief that answers the six basic questions of what, when, who, where, and how.

3. EXPERIMENT ONE DESIGN: NOISE EVALUATION WITHOUT GROUP FEEDBACK:

Participants: A minimum of 30 analysts and postgraduate students will be recruited and receive training in the use of the Client DSRA Team tool.

Procedure: Participants will be briefed on intelligence reports and the developed software. They will then work through a set of intelligence reports (Set A) individually. The report data will be designed to reflect a range of different intelligence report formats and a range of 'interest level'. For each intelligence report, the participants will be required to make a decision about the 'interest level' and rank the report accordingly. These individual judgements will be combined into a collective judgement using the aggregation procedure and then we can estimate how far an individual diverges from group level performance. The aggregation procedure will use inputs from all participants (N=30) to compute the group score for each report.

Note that in this experiment, the individuals will not have contemporaneous access to group ranking. [Also, note also that the aggregation procedure can be repeated in different days. This is due to the fact that we do not expect all data collection to be achieved in a single workshop, since from past experience, when we run experimental studies - including live simulations; they are usually run over several days in order to obtain the numbers necessary for reliable statistical comparisons].

Analysis of Results: This approach will allow us to look at hits (high score given to important reports), and misses (low scores given to important reports), and differences by comparing individual results with the group assessment. Furthermore, judgements made by participants with respect to the 'interest level' of the intelligence report on a Likert scale will enable us to compare individual judgements with each other with respect to the two variables of accuracy and efficiency as defined in the experiment design section. Accuracy is influenced by 'noise' (the chance variability of judgements or inconsistent decisions), and group consensus. Note that in this context although participants are working individually, the analysis can be performed by comparing individual analysis to the group level outcomes.

To facilitate the analysis of results, the DRSA Team

Report	Rank by Analyst 1	Rank by Analyst 2	……..	Rank by Analyst n	Group Rank
R1	1	2		2	2
…					
Rm	5	1		2	3

Table 1: Summary of analysis results

Tool should offer the possibility to generate an Excel spread sheet as in Table 1. The data in Table 1 are for illustration only.

The best way to measure noise is to use the following well-known non-parametric statistics: Kendall's tau, Spearman's rho, and the Unweighted and Weighted Cohen's kappa. See Appendix 1 for more details about these measures. This will be compared with the noise index as described above.

Note: In short, Mark (Polaris) just needs to add to the current version of DRSA Tool: (i) a server/client service to collect/send scoring data between the analysts (client) and the central server; (ii) develop the Aggregation Procedure (See appendix) ; (iii) add the Summary tab and (iv) compile the data (best in Excel files) that will be used (by UoP Team) for evaluation. Aggregation Procedure implements the group decision making as per the EJOR paper (Chakhar et al, 2016).

The output of individual as well as overall group ranking will be used to construct two matrices permitting to measure the agreements levels. The first matrix (see Table 2) will provide agreement level among individual analysts. There will be four different matrices corresponding to the non-parametric statistics (namely Kendall's tau, Spearman's rho, Unweighted Cohen's kappa and Weighted Cohen's kappa). The data in Table 2 are for illustration only.

The second matrix will provide the agreement levels between individual analysts and the aggregated score (see Table 3). The data in Table 3 are for illustration only. As shown in Table 3, the four non-parametric statistics will be used to measure the individual/group agreement level.

These tables will be used by the Noise & Bias Evaluation Tool (NBET) to measure how group and individuals have achieved the target stated at the beginning of this report, that is, a) reach the right answer, and b) within the minimum time. The architecture and data flow within the NBET is shown in Figure 2.

Since the data is collected over intervals of time (or a given number of scored reports), we will have two types of analyses. The first will be conducted once at the end of the experiment using the final results. The second will be based on a time series-like analysis. This will allow us to measure the evolution of individual / individual and individual / group agreement levels across time. The kind of results that we can obtain in the second type of analysis is shown in Figure 3.

Non-parametric statistics	**Analyst 1**	**Analyst 2**	**……..**	**Analyst *n***
Analyst 1	1	0.8		0.3
Analyst 2	0.8	1		-0.4
…				
Analyst *n*	0.3	-0.4		1

Table 2: Agreement levels between pairs of analysts

Analyst	*Kendall's tau*	*Spearman's rho*	*Unweighted Cohen's kappa*	*Weighted Cohen's kappa*
Analyst 1	**0.6**	**0.7**	**0.3**	**0.25**
…..				
Analyst *n*	**0.9**	**0.95**	**0.83**	**0.8**

Table 3: Agreement levels between individual analysts and aggregated score

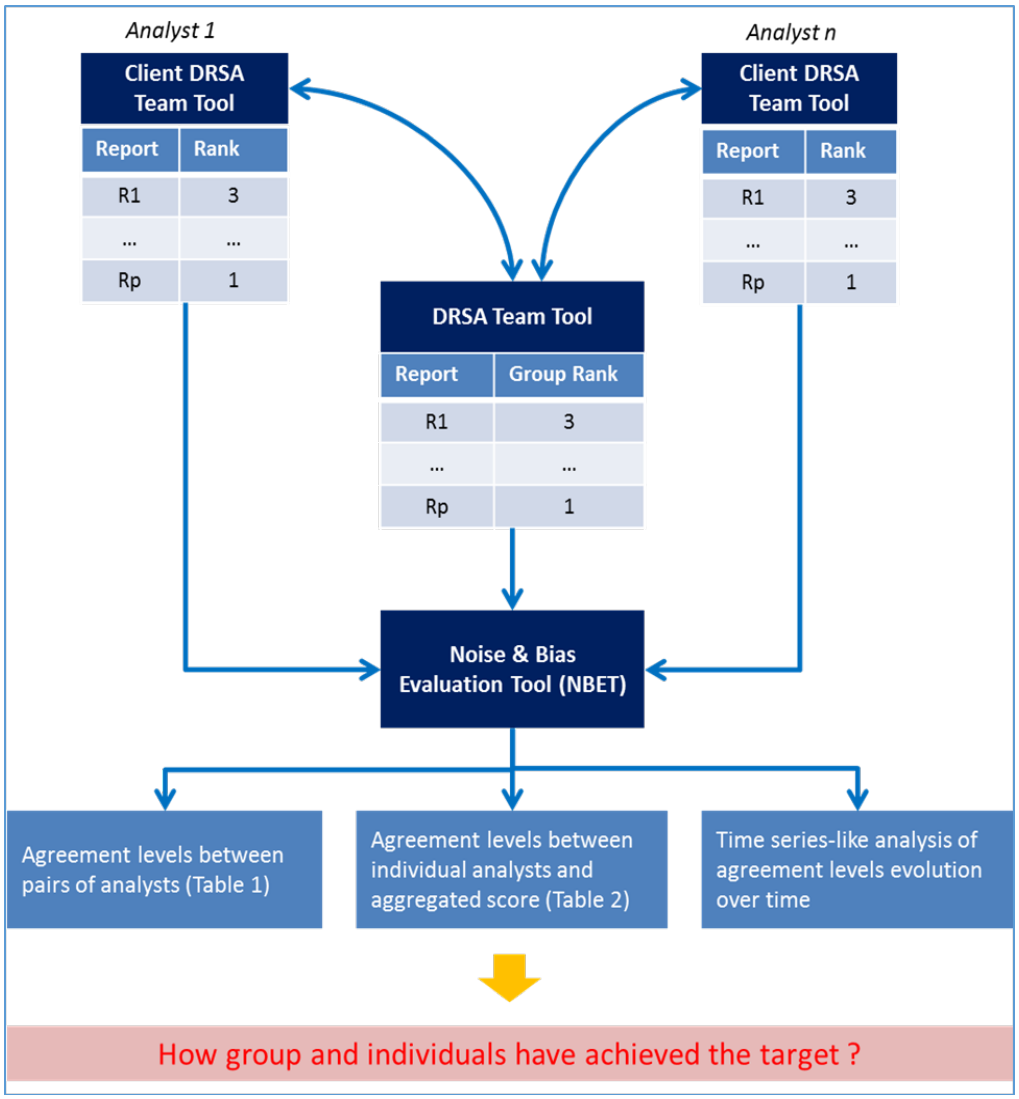

Figure 2: Data flow

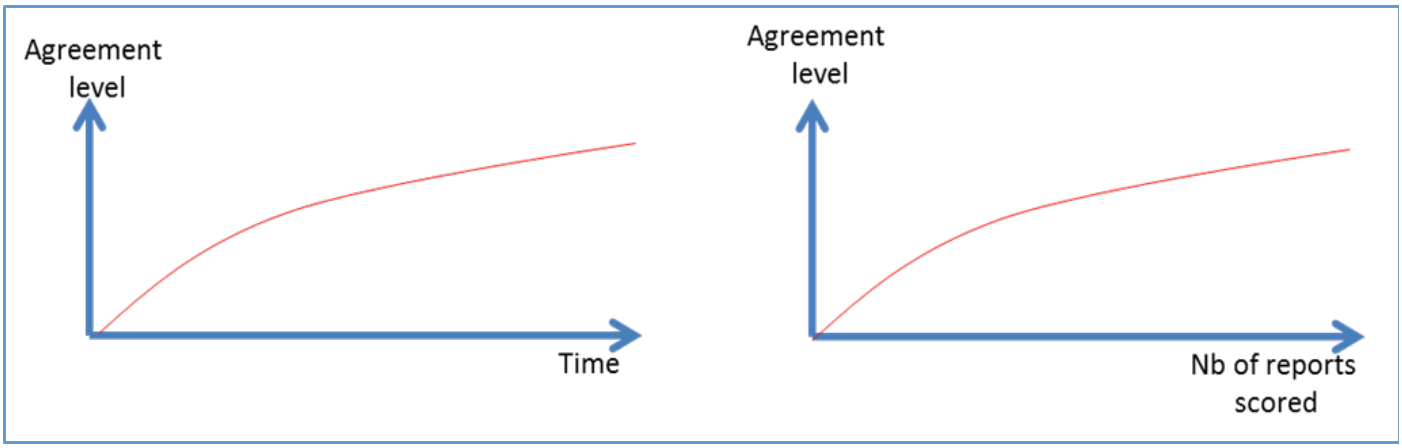

Figure 3: Typical outputs of time series-like analysis results

As indicated earlier, each analyst should submit a final report that summarises his/her scoring process. This report provides answers to the five What, Who, Where, How, and When questions with respect to the critical target event. To construct the overall summary report, we first combine the individual summary reports as in Table 4. Each cell (Analyst i, Question j) in Table 4 contains the response to the question in the corresponding column j as specified by the analysts in the corresponding row i.

*NB: Cl = Confidence Level for each question

The response to each of these questions in the overall summary report will look like as follows:

Question	
Response	**Confidence level**
Response1	**75%**
Response2	**15%**
Response3	**10%**

The first column in this table contains responses given by all the analysts. The second columns indicate the confidence level associated with each response (since the same response may be given by more than one analyst). The confidence level can simply be computed as the percentage of analysts giving the corresponding response. Naturally, the responses to each question will be ordered according to their confidence levels.

4. EXPERIMENT 2 (EVALUATION OF GROUP FEEDBACK AND GROUP BIAS):

In this experiment, we will replicate **Experiment 1** but with additional modification as follows:

We will show participants in real time, through the software, how the group ranking is being done in a dynamic real time environment (an additional column to the existing DRSA column in the scoring screen).

This means that participants will now have access to group level decisions when making their own.

We will also have a group of participants with no access to DRSA or group feedback.

This test will help us to measure the effect of showing the group ranking (group aggregated score) on performance of individuals.

We hypothesise that this additional facility will help them to reach accurate decision (measured by quality of summarised answers to specific questions), reaching high agreement (measured by decrease discrepancy or noise in the judgement scores), and at a faster rate (measured by number of reports per unit time). We should also acknowledge here that it might increase 'false' or inaccurate consensus – and we need to determine a rate for that possible negative outcome, which will be done during the analysis of the reports.

Analysts	*Questions*				
	What?/Cl*	***Who? /Cl***	***Where? /Cl***	***How? /Cl***	***When? /Cl***
Analyst 1					
…..					
Analyst *n*					

Table 4: Summary Final reports

Participants: A minimum of 60 (and that really is a minimum for between groups) different analysts and postgraduate students will be recruited and receive training in the use of the Client DSRA Team Tool.

Design Procedure: Participants will be briefed on intelligence reports and the developed software. They will then run through Set A of reports individually. The report data will be designed to reflect a range of different intelligence report formats and a range of 'interest level'. The individual judgements will be combined into a collective judgement using the aggregation procedure.

In this experiment, third of the individual analysts (n=20) will not have access to global ranking column in the scoring screen, the other third (n=20) will have access to the additional column. The last third (n=20) will have no access to DRSA.

Analysis of Results: This approach will allow us to look at hits, misses and false alarms, and differences by comparing individual results with the aggregated (group) assessment. Furthermore, judgements made by participants with respect to the 'interest level' of the intelligence report on a Likert scale will enable us to compare individual judgements with each other with respect to the two variables of accuracy and efficiency.

As part of the analysis, we will construct randomly subgroups of n=5 to represent a typical intelligence cell. In order to ensure more accurate results, we will randomise the groups 3 times which will lead to 9 different configurations of groups for those with feedback information on group ranking and another 9 different configurations for those without such information. Therefore, this will be fed back to the participants (i.e. from the live decisions being made. In other words, participants will have an additional column is their tool showing (live) the group ranking as computed by the aggregation procedure.

The analysis conducted for Experiment 1(Tables 1&2) will also be conducted for Experiment 2 for each group. Form both experiments, we hypothesize the following:

1. That a high agreement on ranking on scoring will lead to a high agreement to final summary report (i.e. that IAs are rationale in transforming numeric score into narrative report).
2. That the group-based results will be of higher quality than those from IA (i.e. majority of the group decisions are better than some individuals), and that the majority of ISs with DRSA access will perform better than those without.
3. That those IAs with information feedback about 'group behaviour' will do the exercise more efficiently.
4. That by using our tool with information about group behaviour, the participants will be both effective and efficient in the decision they take.

5. ADDITIONAL ANALYSIS:

For both experiments, the summary report will be compiled at the end of each experiment and this information will be used to measure 'accuracy' of each individual. In addition, within Experiment 2, this type of analysis will allow us to verify the effect of 'group bias' (following the crowd) on individual's performance.

The ranking results of the report vary across time: The ranking of the reports will provide us with an additional opportunity to measure the level of agreement across time with respect to the number of reports.

The Team bias for a given report can be measured by aggregating individual biases and also by measuring the difference between the group score of the original and altered report. An overall group bias can then be measured as the rate of altered reports ranked differently by the team.

APPENDIX 1

NON-PARAMETRIC STATISTICS USED FOR THE ANALYSIS

The following well-known non-parametric statistics will be used to measure the agreement levels:

- *Kendall's tau*. Kendall's tau lies in the range [-1,1]. If the agreement between the two rankings is perfect (i.e. the two rankings are the same) it is 1. If the disagreement between the two rankings is perfect (i.e. one ranking is the reverse of the other) it is -1. If two rankings are independent, then we would expect it to be approximately zero.
- *Spearman's rho* . Spearman's rho is in the range [-1,1]. A positive Spearman correlation coefficient indicates that both rankings vary in the same direction. A negative Spearman rho coefficient indicates a monotone decreasing relation between the two rankings. A Spearman rho coefficient of zero indicates that there is no tendency between the two rankings.
- *Cohen's kappa*. There are two ways of calculating Cohen's kappa: unweighted and weighted. The weighted kappa is more appropriate for variables having more than two categories. In both cases, the value of Cohen's kappa lies in [0,1]. Conventionally, a kappa of <0.2 is considered poor agreement, 0.21--0.4 fair, 0.41--0.6 moderate, 0.61--0.8 strong, and more than 0.8 a near complete agreement.

We are familiar with these statistics. If we get an Excel file with the scorings of the reports by all analysts, we can easily compute these statistics.

It is important to mention that these statistics accept ordinal data and can deal with ties. Furthermore, they take into account the number of levels between any two compared scores.

We may also use two additional statistics, namely Kendall's W and/or Fleiss's kappa. The Kendall's W (also known as Kendall's coefficient of concordance) is used for assessing agreement among multiple rankings. The Fless's kappa is an extension of Cohen's kappa to evaluate concordance or agreements between multiple rankings. The Kendall's W and Fless's kappa are devoted to compare at least three different rankings. In addition, both of them accept ordinal data and can deal with ties and also take into account the number of levels between compared scores.

APPENDIX 2

DRSA

The Dominance-based Rough Set Approach DRSA[6] is a well-known multicriteria classification method. As shown in Figure 5, the working mechanism of the DRSA is a typical machine learning manner and often categorized as 'preference learning' method. The DRSA takes as input a subset of scored reports as learning examples and generates a set of 'if..., then...' decision rules. First, the analysts should score a collection of reports with respect to the considered five level scale. The latter defines five preference-ordered classes from Cl_1 to Cl_5 with an increasing importance level.

Then, each of these classes is represented in terms of its lower and upper approximations. The lower approximation of class Cl_t contains all the reports that certainty belong to class Cl_t while the upper approximation of class Cl_t contains all the reports that may belong to Cl_t. The difference between the upper and lower approximations, called boundary (or doubtful) region, groups all reports that can be ruled neither in nor out as members of class Cl_t. When the approximation is perfect, the lower and upper approximations are equal, the boundary will be empty.

The obtained approximations are then used to infer a set of 'if..., then' decision rules. Three types of decision rules may be considered in DRSA: (i) certain rules generated from lower approximations; (ii) possible rules generated from upper approximations; and (iii) approximate rules generated from boundary regions. Only certain decision rules are considered here. Their general structures are as follows:

IF *condition(s)* THEN *Importance Level = At Least* Cl_t

IF *condition(s)* THEN *Importance Level = At Most* Cl_t

The condition part specifies values assumed by one or more condition attributes and the decision part specifies an assignment to one or more decision classes. The decision part of certain rule takes the form of an assignment to at most or at least class unions. The decision part of a possible rule is a union of several decision classes.

The obtained decision rule can finally be used to classify the unseen reports.

Figure 4: Principle of DRSA

6 See Greco, S., Matarazzo, B., Słowiński, R.: Rough sets theory for multi-criteria decision analysis. European Journal of Operational Research, 129, 1 (2001) 1–47.

APPENDIX 3

AGGREGATION PROCEDURE

The DRSA method has been designed for single decision makers. The aggregation procedure extends the DRAS to work with multiple decision makers (Chakhar et al, 2016). Before applying the aggregation procedure, the DRSA method is used to approximate the input data provided by each analyst. Then, the aggregation procedure is used to construct a collective decision table[7], which is later fed to DRSA to generate a set of collective decision rules. The collective decision rules that are used to assign the overall score to all the reports.

The basic idea of this procedure is to use the outputs of individual classifications to assign to each report x an assignment interval [l(x),u(x)] where l(x) and u(x) are respectively the lower and upper classes to which report x can be assigned, and then some simple rules are used to reduce the assignment interval I(x) into a single element representing the final and overall score of the report. The computation of overall score relies on the majority principle and veto effect.

The contribution of each analyst to the collective score is measured by the quality of input data provided by the analyst. The use of input data to deduce the contribution of the analysts is generally more objective than the other weighting techniques.

The aggregation procedure is designed to work for two or more analysts. The analysts may share or not the evaluation criteria (e.g. keywords, places, people, etc.), but they need however to use the same scoring scale.

Reference:

Chakhar, S., Ishizaka, A., Labib, A., Saad, I., Dominance-based Rough Set Approach for Group Decisions, European Journal of Operational Research, Vol 251, pp206-224, 2016.

7 The decision table is a matrix where the rows stand for reports and columns to criteria (such as keyword, people, etc.). The last column of the decision table represents the scores as specified by the analysts.

For more information on CREST
and other CREST resources, visit
www.crestresearch.ac.uk

CREST

CENTRE FOR RESEARCH AND
EVIDENCE ON SECURITY THREATS

20-018-01