Leggatt, A. & McGuinness, B. (2006) Factors influencing information trust and distrust in a sensemaking task. *11th International Command and Control Research and Technology Symposium*, 26-28 Sept 2006, Cambridge, UK (DoD Command & Control

Factors Influencing Information Trust and Distrust in a Sensemaking Task

Andrew Leggatt and Barry McGuinness
BAE Systems, UK

ABSTRACT

Information trust refers to a user's willingness to accept a given piece of information into a decision-making process when the use of "bad" information could be a critical mistake. This paper reports on a DoD-funded experiment into teams' patterns of information trust in the context of a military sensemaking task. It follows on from a preliminary experiment conducted in 2005, in which it was found that the users' awareness of a message's source (and any assumptions or biases associated with that) had an overriding influence on their decision to trust or not trust the information it contained (McGuinness & Leggatt, 2006). For this second experiment, pairs of subjects undertook a collaborative task of compiling an accurate intelligence picture during a simulated coalition engagement with enemy forces. Awareness of information sources was manipulated by running the task in one condition with "anonymous" sources only (i.e., message sources were not revealed). In addition, the insertion of a (bogus) network security alert was used as a way to prompt greater attention to information quality. It was found that teams actually made better trust/distrust judgements with anonymous sources and when given a network alert, presumably because they paid more attention to the information itself.

INTRODUCTION

Enormous advantages are anticipated in the ability of networked information technology to reduce the fog of war. At the same time, however, command and control is becoming increasingly dependent upon information, and upon the confidence people have in that information. This dependence is widely regarded as a potential Achilles' heel in the network-centric war-fighting organization, as information systems and networks are prone to a variety of potential weaknesses which can affect the trustworthiness of the information available to decision makers. Moreover, human uncertainty regarding the trustworthiness of information can lead to a lack of confidence that inhibits proactive decision-making and action (Alberts et al, 2001). Knowing what information to trust, and when to distrust information, are therefore important human issues.

Information trust can be defined as a willingness to accept a given piece of information into one's decision-making processes (where the use of invalid or unreliable information could lead to a critical decision error), on the grounds that it is believed to be sufficiently valid and reliable (McGuinness, 2004). Information that is not deemed sufficiently valid or reliable may be rejected and discarded; this unwillingness to accept such information is termed information distrust. There are then two distinct and potentially critical types of error that can occur in relation to trusting information:

- *Type 1 error*: Accepting and using information that is not in fact trustworthy (i.e. "mistrust").

- *Type 2 error*: Rejecting and discarding information that is, in fact, trustworthy (for which we have coined the term "misdistrust").

To the extent that there is any threat to information quality, there is a need for individuals to be able to make appropriate information trust/distrust judgements in order to avoid these errors. As yet, however, there has been little research of direct relevance. The causes and effects of information trust and errors of information trust are largely unexplored. For instance, in information-rich, networked military environments, just how good are information users at judging what information to trust and what information to distrust? What are the relative risks and consequences of trusting the "wrong" information and distrusting the "right" information? What factors makes users more prone to such errors? How can we can improve users' judgements of information and thereby minimise the risk of erroneous trust or distrust?

Information Trust in Network Centric Operations (Phase 1) is a 12-month DoD-funded project undertaken by the Advanced Technology Centre (ATC) of BAE Systems in the UK. [1] The overall aim of this research is "to investigate and understand the causes and effects of appropriate and inappropriate information trust and distrust in the context of network-centric operations." Two experiments were conducted in Phase 1 (2005-06) with the assistance of the UK's Defence Academy. The second experiment is the focus of this paper, but it is necessary to begin with a summary of the methods and findings of the first experiment.

EXPERIMENT 1

Experiment 1 investigated the influence of situational awareness/understanding on the appropriateness of information trust responses. Twenty-two British Army majors studying at the Defence Academy, Shrivenham, performed an individual sensemaking task designed to elicit multiple information trust/distrust discriminations. During the task, which was set within a coalition operation based in Africa, each received a flow of intelligence reports/messages from multiple virtual sources on the status, position and movement of enemy forces. The content of these messages was automatically translated into appropriate visual icons on an electronic map. The subjects' task objective was to read and make sense of the incoming information in order to assess the enemy's course of action, though making sure to filter out any untrustworthy information in the process. In fact, 25% of all messages had some kind of information quality 'defect', which could be inaccuracy, incorrectness, incompleteness, untimeliness, irrelevance, or inconsistency. The subjects' acceptance or rejection of each message and their confidence rating for each message were analysed to provide various measures of information trust.

Two key factors were manipulated: (1) the subjects were given either a high or low prior understanding of the situation at the start of the task; (2) at one point the subjects were presented an information network alert informing them that a breach of the network had occurred and that information quality may have been compromised. It was hypothesised that

[1] DoD contract number W74V8H-05-P-0288.

better understanding of either (a) the operational situation or (b) the network status would facilitate better judgements of information trustworthiness. In fact, these interventions had little effect on the information trust data. This was found to be due to an overriding effect of the subjects' awareness of the information sources, and the biases and assumptions associated with that.

EXPERIMENT 2 – AIMS

When the results of Experiment 1 were presented to a group of subject-matter experts at the Defence Academy, the idea was raised that eliminating the prior assumptions associated with information sources could actually improve trust/distrust judgement. It had seemed from Experiment 1 that the more a source is trusted on the basis of *a priori* assumptions, the less attention users paid to the content of information from that source. Eliminating source awareness, then, might actually lead to an improvement in judgement. It was therefore decided to compare responses to information with known sources versus responses to information with unknown sources, i.e. using "anonymous" sources in one condition. The experiment was designed to test the seemingly counter-intuitive hypothesis that ***awareness of an information item's source can actually impair judgement as to the real trustworthiness of information.***

Experiment 2 repeated the methodology of Experiment 1, though with a few significant modifications.

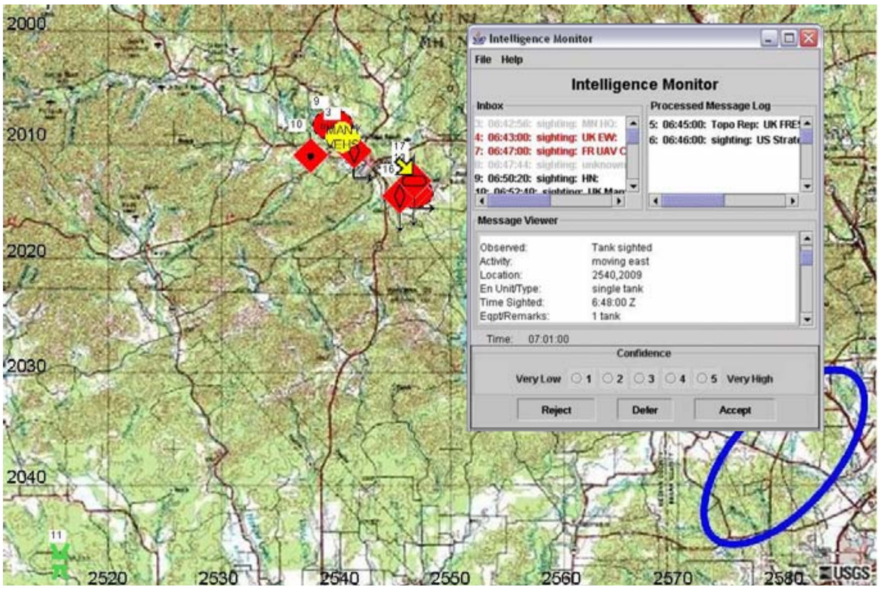
METHODOLOGY

Task

The task was set within a fictitious future coalition operation based in Africa, where rebel forces were preparing to attack the capital city of a former French colony, 'Kumbiba'. A French-led multinational HQ was supported by a combination of French, British, American and host-nation armed forces. The subjects were theoretically located within the intelligence cell of the British element's HQ. In the course of the operation, they received a flow of electronic intelligence reports/messages from multiple sources on the status, position and movement of enemy forces. The content of these messages was automatically translated into appropriate visual icons on an electronic map (Figure 1). There were 54 messages per scenario. All were delivered to the subject electronically using a system akin to email. The subjects could read and process incoming messages using a tool called the Intelligence Monitor (Figure 2), the user interface of which consisted of the following components:

- Inbox (showing received items)
- Message Viewer (where message text could be read)
- Processed Message Log (storing messages that have been accepted)
- Five-option rating scale to give confidence ratings
- Three buttons marked **ACCEPT**, **DEFER** and **REJECT**

Messages arrived in the Inbox at pseudo-random intervals of approximately two messages per minute.



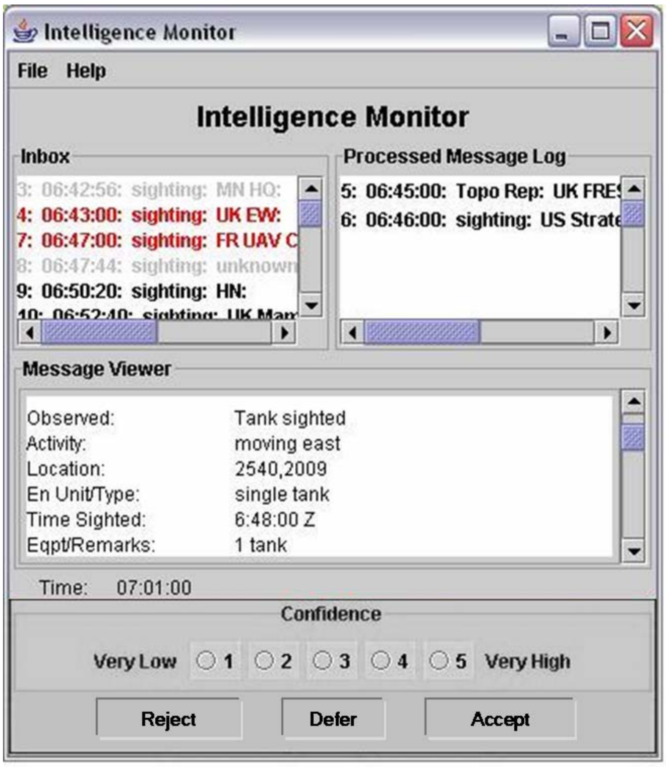
3: 06:42:56: sighting: MN HQ:
4: 06:43:00: sighting: UK EW:
7: 06:47:00: sighting: FR UAV C
8: 06:47:44: sighting: unknown
9: 06:50:20: sighting: HN:
10: 06:52:40: sighting: UK Man

5: 06:45:00: Topo Rep: UK FRES
6: 06:46:00: sighting: US Strate

Observed: Tank sighted
Activity: moving east
Location: 2540,2009
En Unit/Type: single tank
Time Sighted: 6:48:00 Z
Eqpt/Remarks: 1 tank

Time: 07:01:00

Figure 1: Screenshot of the electronic map and message software used by the subjects



3: 06:42:56: sighting: MN HQ:
4: 06:43:00: sighting: UK EW:
7: 06:47:00: sighting: FR UAV C
8: 06:47:44: sighting: unknown
9: 06:50:20: sighting: HN:
10: 06:52:40: sighting: UK Man

5: 06:45:00: Topo Rep: UK FRES
6: 06:46:00: sighting: US Strate

Observed:	Tank sighted
Activity:	moving east
Location:	2540,2009
En Unit/Type:	single tank
Time Sighted:	6:48:00 Z
Eqpt/Remarks:	1 tank

Time: 07:01:00

.

Figure 2: Screenshot of the subjects' electronic message reader

The subjects were instructed to review and process each message as soon as possible after it was received, using the following procedure:

1. Select an item in the Inbox
2. Read the message that appears in the Message Viewer
3. Give the message a reliability confidence rating of 1–5 (1 = very low, 5 = very high).
4. Select either **ACCEPT** to retain that message as part of the current intelligence picture, or **REJECT** to remove it. Alternatively, press **DEFER** to leave the item in the Inbox and return to it at a later time. Then repeat step 1.

Performing in two-person teams, the subjects were asked to fully collaborate on the message-processing task such that the sensemaking and information trust results would reflect their joint decisions and judgements. They were also told that the task was not over until they had fully processed (accepted or rejected) *all* messages.

The subjects rehearsed and then performed the task in a networked laboratory/workshop facility at the Defence Academy in Shrivenham (Figure 3).

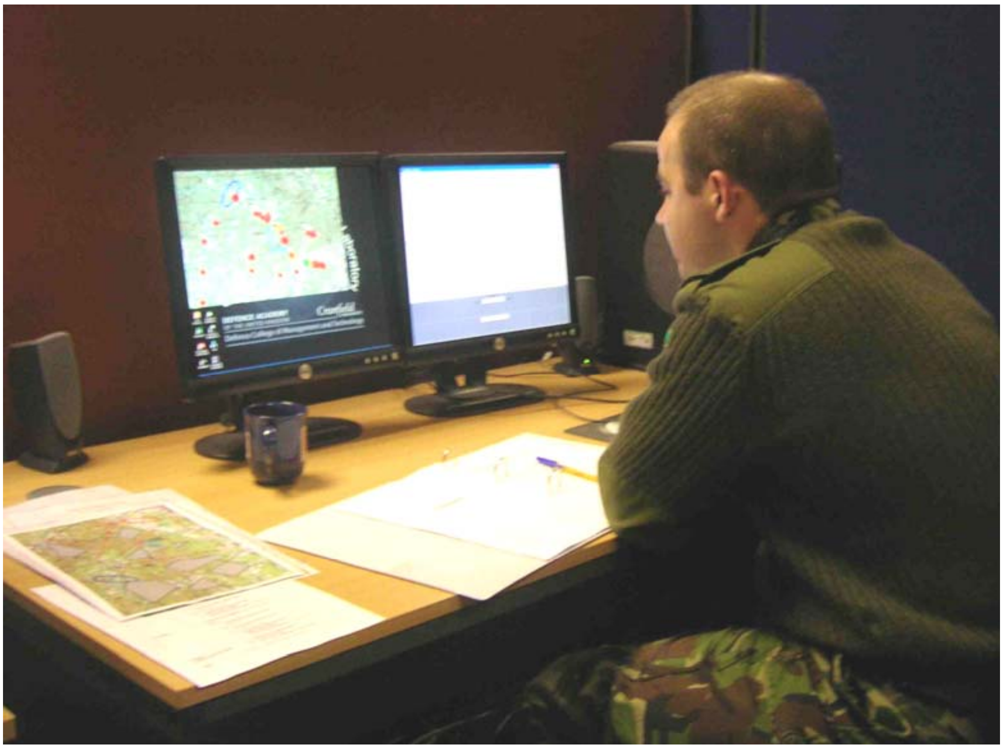


Figure 3: A subject undergoing training in the message processing task

Independent variables

The following factors were manipulated:

- Awareness of Information Source
- Occurrence of a Network Alert
- Information Quality

In Experiment 1, a variety of intelligence sources had been invented chiefly to provide a realistic context for the task. In Experiment 2, source itself was treated as a factor of interest. The subjects' awareness of source identity was manipulated by either disclosing or not disclosing the source label for each message. Thus, source identification was varied within subjects to give two conditions,

- Known (disclosed)
- Anonymous (not disclosed).

The subjects in Experiment 1 had appeared to favour some source nationalities over others, and they appeared to favour technical over human intelligence. It was therefore decided to vary the source *nation of origin* and the *intelligence type*. There were three nations of origin: *UK*, *USA* and "*Kumbiba*" (the African host nation). There were two intelligence types: *technical* (sensor-derived intelligence) and *HUMINT* (human-derived intelligence). By combining nation of origin (x3) and type of intelligence (x2), there were six possible source identifications. In the Known Source condition, the sources were ascribed as "UK – Technical", "USA – HUMINT", and so on. In the Anonymous Source condition, the sources were identified only by a code-letter (A–F), as shown:

'Known' identification	**'Anonymous' identification**
• UK Technical	• D
• UK HUMINT	• B
• USA Technical	• A
• USA HUMINT	• F
• Kumbiba Technical	• C
• Kumbiba HUMINT	• E

The subjects were not, of course, told to which sources these letters corresponded.

Occurrence of a Network Alert had two conditions:

- Alert Given
- No Alert Given

During one of their two runs the subject was handed an alert message informing them that "a breach of the intelligence network" had been detected and that "information quality may be temporarily compromised." (In fact, the ratio of good information to flawed information remained the same at 3:1.) The Alert was presented $^{1}/_{3}$ of the way into the task; an "Alert Cancelled" message followed at the $^{2}/_{3}$ point. In the other run, no such alert was given.

Information Quality varied between:

- High Quality (correct information)
- Low Quality (incorrect information)

As in Experiment 1, 25 per cent of all messages were of low quality. Unlike those used in Experiment 1, however, the low quality messages used in Experiment 2 had only one type of information defect, namely incorrectness. The other types of low information quality used in Experiment 1, such as untimeliness and imprecision, were not included. Hence, a given message was either correct or incorrect as an objective description of some part of the situation.

Design

Following a repeated-measures design, all subjects performed the task twice, using two similar but non-identical versions of the scenario. One run was performed under the Known Source condition, while the other was under the Anonymous Source condition. Likewise, one run contained a network alert while the other did not. The running order of source identification and alert conditions across subjects, and the use of scenarios, was balanced as shown so as to minimise order effects.

Subjects

The subject set consisted of 9 Army majors, all British males currently studying at the Defence Academy, Shrivenham, with 8-14 years in service, a further 2 members of staff from the Defence Academy (with military backgrounds), and 8 civilians from the BAE Systems ATC. Their ages ranged from 21 years to 51 years, with an average of 31.6 years.

Dependent variables

The performance attribute of interest was the appropriateness of subjects' trust and distrust of the messages received. These were operationalised through the following dependent variables:

- The teams' subjective confidence in each message (self-ratings on 5-point scale)
- The teams' objective response error rates, i.e.
    - Type 1 error rate (proportion of incorrect messages accepted)
    - Type 2 error rate (proportion of correct messages rejected)

Aside from providing error rates, the ACCEPT/REJECT response data also lend themselves to analysis using the framework of Signal Detection Theory (SDT). Specifically, high-quality information items may be treated as 'signals' to be discriminated from low-quality items, regarded as 'noise' or non-signals, as shown in Table 1:

Table 1: Contingency table of stimulus-response outcomes

		Response type	
		ACCEPT	REJECT
Information quality	HIGH	**Hit**	**Miss** (type 2 error)
	LOW	**False Alarm** (type 1 error)	**Correct Rejection**

SDT can provide two very useful statistics that are computed from the proportion of hits and false alarms obtained in an experiment, namely:

- *d'*, the degree to which 'signal' and 'non-signal' stimuli are correctly discriminated.
- *ß*, the degree of response bias (if any) in favour of signals or non-signals.

In the present experiment, *d'* represents the subjects' ability to differentiate good from bad information while *ß* represents the subjects' tendency to either trust (accept) or distrust (reject) potentially unreliable information.

A variety of other measures were taken to assist the interpretation of the results. These included:

- Understanding of the enemy course of action (multiple-choice questions)
- Expected reliability of each source (pre-trial ratings on 5-point scale)
- Perceived reliability of each source (post-trial ratings on 5-point scale)
- Personal trust/distrust attitudes (14-item questionnaire)
- Personal background details

RESULTS

Source reliability ratings

Ratings for source reliability, both pre-trial (expected reliability) and post-trial (perceived reliability), are shown in Figure 4.

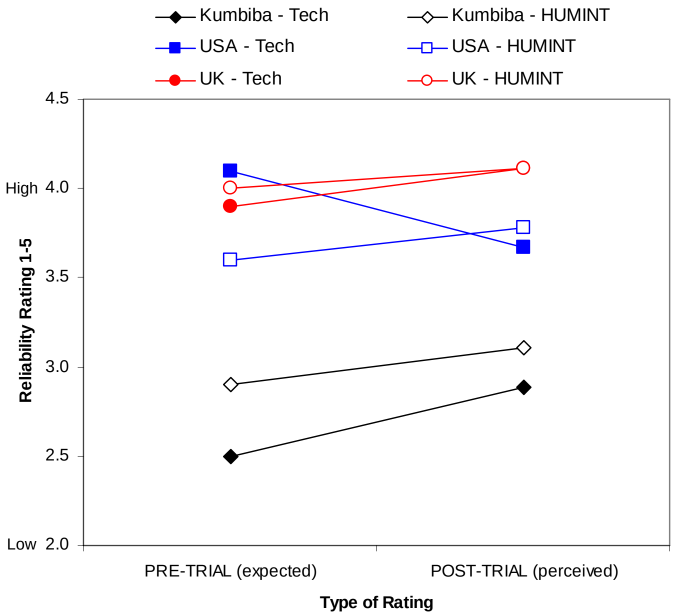


Figure 4: Subjects' pre- and post-trial ratings of source reliability

As can be seen, there was a lot less confidence in the African host nation sources than in the UK or USA sources – though in actuality all sources were equally reliable at 75%. This suggests, consistent with our Experiment 1 findings, that the subjects had prior assumptions about the sources which may have affected their perceptions of information trustworthiness.

Information confidence ratings

Turning now to the teams' ratings of confidence in individual information items (Figure 5), we find a significant interaction between the two main factors, Awareness of Source and Occurrence of Network Alert ($F_{1,11} = 6.67$; $p < 0.01$).

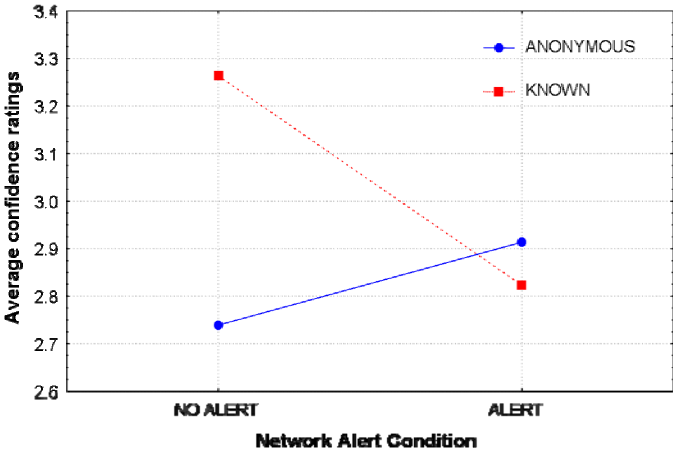


Figure 5: Mean ratings of confidence in information items under each condition

On those runs in which no network alert was given, the average rating of confidence in information was half a point higher with known sources than with anonymous sources. On those runs in which a network alert *was* given, however, confidence ratings with known sources reduced to the same level as those with anonymous sources. The pattern of results suggests that *either* not being aware of source identity *or* being alerted to a potential problem with network reliability is sufficient to reduce confidence in information (though there is no additive effect of both together).

Error rates

The average error rate for all ACCEPT/REJECT responses was 0.27. Type 2 errors were made twice as often as type 1 errors. The overall type 1 error rate (i.e. the false alarm rate, the proportion of incorrect messages that were accepted) was 0.71, while the overall type 2 error rate (i.e. the miss rate, the proportion of correct messages that were rejected) was 0.18.

There was a significant main effect of Awareness of Source on both of the error rates. The average false alarm rate was 0.53 when sources were known, but rose to 0.71 when sources were anonymous ($F_{1,48}$=8.29; p<.0047). In other words, type 1 errors were less likely when information sources were identified. The average miss rate, in contrast, was 0.25 when sources were known, but *fell* to 0.11 when sources were anonymous ($F_{1,48} = 6.53$, p <.0100). In other words, type 2 errors were *more* likely when sources were identified.

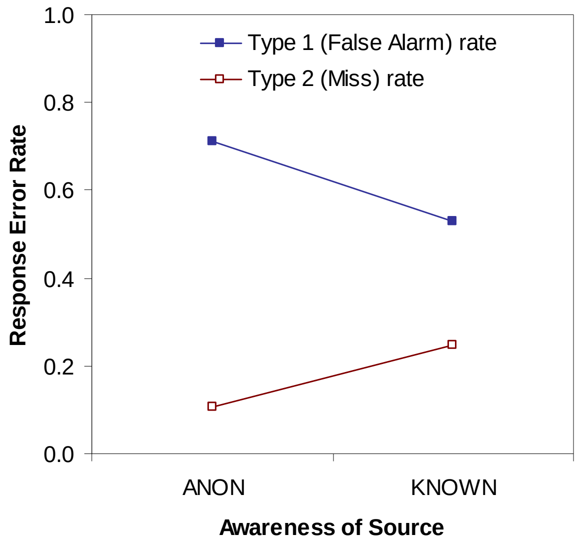


Figure 6: Type 1 and type 2 error rates with 'known' and 'anonymous' sources

Interestingly, the miss rates were unaffected by Occurrence of Network Alert (the average being 0.19 in both conditions), but there was a significant effect of the network alert upon false alarms. In fact, the occurrence of an alert *reduced* the average type 1 error rate from 0.71 to 0.53 ($F_{1,48}$= 4.98; p<.0276).

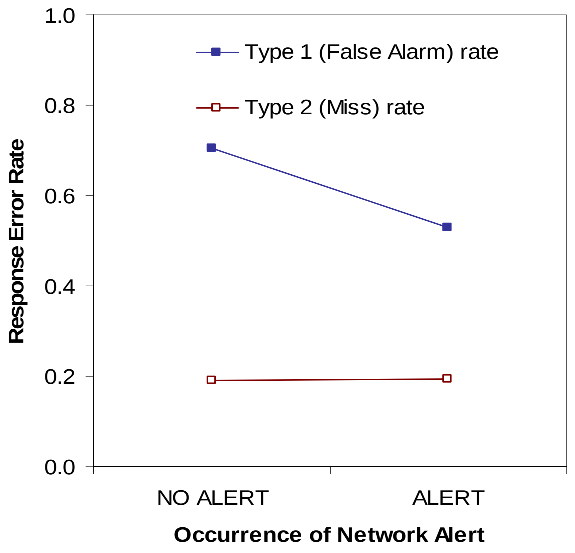


Figure 7: Type 1 and type 2 error rates with and without a Network Alert given

Signal detection analysis

We now examine the ACCEPT/REJECT responses using the statistical and graphical methods of Signal Detection Theory. Figure 8 shows a response operating characteristic (ROC) plotting hit rates against false alarm rates under the combined conditions of Awareness of Source and Occurrence of Network Alert. The statistics for sensitivity (*d'*) and response bias (*ß*) associated with each of these are summarised in the table below.

Table 2: SDT statistics per condition

	CONDITION			
	Known Sources		Anonymous Sources	
	No Alert	Alert	No Alert	Alert
d'	0.25	0.77	0.24	1.16
ß	0.85	0.85	0.79	0.35

To begin with response bias (*ß*): Given that good information was three times more numerous than bad information in this experiment, the optimal value of *ß* was found to be 0.33 (as represented in the ROC chart by the isobar denoted $ß_{opt}$).[2] As we can see, only in one condition did the subjects demonstrate this optimum: this was with *anonymous* sources and with an *alert* given. In all other cases response bias was distinctly cautious or "conservative", leading to fewer false alarms but more misses. The non-disclosure of source identity and the presentation of a network alert together appear to have had an additive positive effect on *ß*.

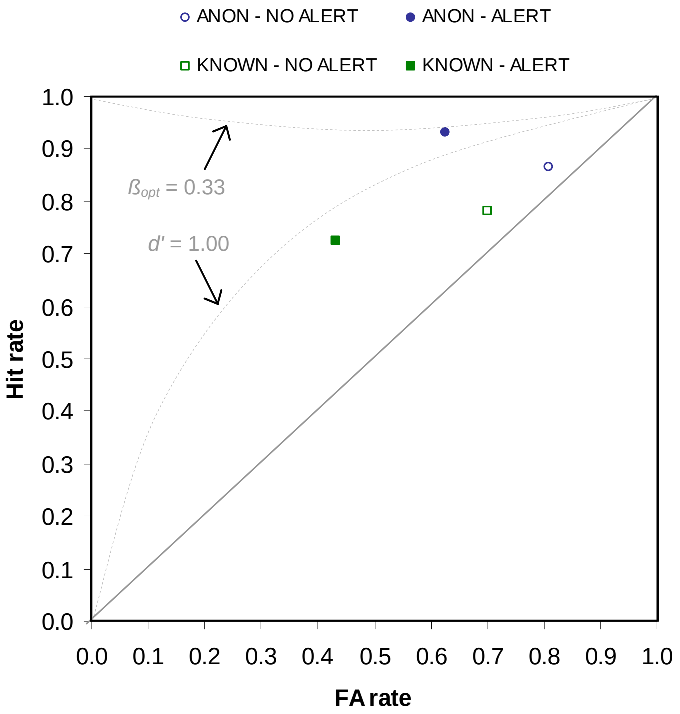


Figure 8: ROC chart based on subjects' ACCEPT / REJECT responses

The ROC chart also shows an isobar for *d'* = 1. A high *d'* value (of the order of 3 or 4) represents a good ability to discriminate signal from noise (or in this experiment, high quality information from low quality information). A *d'* value of <1 is considered low and indicates poor discrimination.[3] As can be seen, only in one case (anonymous sources, alert given) was the average *d'* value greater than 1. The greatest effect on *d'* was in fact the presence of the

[2] The optimal value of *ß* is the ratio of the probability of a non-signal to the probability of a signal (Wickens, 2002, p.33).

[3] A *d'* value of 0 indicates *no* systematic discrimination between signals and non-signals.

network alert ($F_{1,48}$=5.56; p<.02). Again, there appears to have been an additive effect of the two experimental manipulations in combination: the teams were best at distinguishing correct from incorrect messages when (a) they did not know the source of any message and (b) they had been alerted to a possible network security problem.

This effect of the network alert can be appreciated with a closer examination of the *d'* data. As can be seen in Figure 9, the teams reacted differently to messages from different sources (when source was known), depending upon whether or not there was an alert given. Without the alert, the teams were seemingly incapable of accurately discriminating correct vs. incorrect messages from the two most "trusted" national sources, UK and USA. Their discrimination of messages from the African source, however, was >1. Contrast this pattern with that obtained when an alert was given. In this case, their ability to discriminate messages from UK and especially US sources improved, while there was a slight drop in discrimination with respect to messages from "Kumbiba". There were no differences, however, between technical and HUMINT intelligence types.

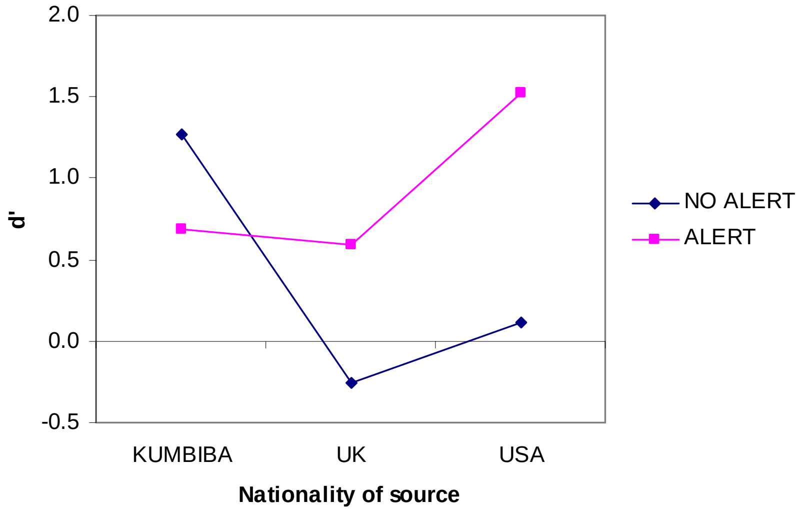


Figure 9: Comparison of d' in relation to known sources

DISCUSSION

The results of Experiment 2 supported the hypothesis with a small qualification. When sources were made "anonymous" there was indeed a *greater* ability to discriminate between good and bad information. This improved discrimination, however, only occurred when the subjects were also alerted to a possible disruption of network security and hence information quality.

These results can be understood with reference to the interpretations of Experiment 1 offered by a group of military subject matter experts (SMEs), as discussed in our previous paper (McGuinness & Leggatt, 2006). The SMEs suggested that information items can be assessed in terms of three categories of perceived trustworthiness, each category entailing a different response:

1. If the information's source is deemed "probably reliable", the user will automatically accept the information.
2. If the information's source is deemed "probably unreliable", the user will immediately set the information aside.
3. If the information's source reliability is unknown or ambiguous, the user will consider the content of the information itself (does it fit the expected pattern, or does it contradict something known for certain?) and seeking further evidence that can confirm or disconfirm it (e.g., monitoring how the situation unfolds, or asking around, "How reliable is that source?" or "Could the enemy really have troops over there?"). This is obviously time-consuming and best avoided if possible.

Note that an item does not have to be assessed as *definitely* reliable for it to be given immediate use; *probably* reliable is sufficient for practical purposes. This of course entails some risk of type 1 errors, i.e. using false or flawed information. The SMEs explained that in tasks such as this where speed is usually of the essence, this risk has to be accepted. Indeed, it seems that "when the heat is on" under time pressure, increasing amounts of ambiguous information will be readily accepted as there is less time or mental capacity available to consider trustworthiness issues. (We could predict, therefore, that type 1 error rates in an information trust/distrust judgement task will be found to positively correlate with mental workload.)

Applying this interpretation to Experiment 2, the results again support the notion that unless there is an obvious reason for not doing so, identity of source is habitually used as the basis for quickly assessing the probable trustworthiness of information. When there is no obvious and immediate sense of source reliability, however, the user will pay more attention to the content of the information itself. More broadly, whenever there is any perceived reason to not take the information at face value – e.g., the source is unknown, or the information network is potentially insecure – then more attention is paid to the information content. As a result, bad information is more readily detected, resulting in fewer type 1 errors, and good information from what would otherwise be non-trusted sources becomes more apparent, resulting in fewer type 2 errors.

The potential to improve performance by alerting attention to information quality rather than just source identity may have practical implications. It is possible that users can be educated to assess information more mindfully by being more aware of the risks of type 1 and type 2 errors and the effects of bias based on prior assumptions associated with source identity. It is also possible that training with feedback could enable users to improve their overall sensitivity to information quality – hence, raising d'. The artificiality of the bogus "network alert" used in this experiment does not preclude the investigation of some more naturalistic means of alerting users to pay attention to message content when errors of information trust or distrust could prove critical.

By better understanding information trust in the NCO context we can seek ways of reducing or minimising the risk of people trusting bad information and disregarding good information. Such solutions might include improved processes for information sharing and usage, or improved personnel selection procedures which take into account individual differences in information trust behaviour, or improved training for information users with regard to assessing and managing information quality. In designing and supporting network-centric organizations, our goal should not only be to ensure that the information available to warfighters is trustworthy as far as possible, but, as we have shown here, should also include identifying and finding ways to minimise the risks of human error. Further research is therefore recommended to look at other variables affecting the appropriateness of information trust judgements and the management of trust errors.

REFERENCES

Alberts, D.S., Gartska, J.J., Hayes, R.E. & Signori, D.A. (2001) *Understanding Information Age Warfare*. Washington DC: CCRP Publication Series.

McGuinness, B. (2004). *Trust, Mistrust and Distrust of Information*. White paper produced for Evidence Based Research, Inc., on behalf of the Office of the Assistant Secretary of Defense for Networks and Information Integration (OASD/NII). BAE SYSTEMS Advanced Technology Centre, Bristol, U.K.

McGuinness, B. & Leggatt, A. (2006) Information trust and distrust in a sensemaking task. *Command and Control Research and Technology Symposium*, June 2006, San Diego (DoD Command & Control Research Program).