DETECTING DECEPTION WITHIN A PROBABILISTIC MODELLING FRAMEWORK

Ken R. McNaught

Cranfield University, Decision Analysis & Risk Modelling Laboratory, Dept of Informatics and Systems Engineering,
Defence Academy, Shrivenham, Swindon, UK.

K.R.McNaught@cranfield.ac.uk

ABSTRACT

In this paper, we consider a number of analytic approaches to identifying or accounting for possible deception tactics being employed by an adversary. These are equally applicable to military or civil intelligence, or even law enforcement. As well as examining the Analysis of Competing Hypotheses (ACH) methodology, employed by some intelligence agencies, we focus on the contribution of methods for reasoning under uncertainty, particularly Bayesian networks (BNs). We also discuss the combination of these approaches as suggested by other authors. It is shown that the incorporation of negative evidence in addition to positive observations improves the performance of the BNs.

Keywords: deception, adversarial reasoning, military intelligence, intelligence analysis

1. INTRODUCTION

Deception is an integral part of human adversarial interaction. However, despite its well-proven and widely accepted value, including many historical accounts, there is relatively little scientific literature regarding its worth. In this paper, we view deception in the context of uncertain reasoning and so adopt a probabilistic modelling approach. This is aimed particularly at detecting possible deceptions.

Although deception is inextricably linked to psychology, the focus of this paper is on modelling issues rather than psychological ones. In particular, we focus on analytic approaches to detecting deception. Such approaches are applicable in a wide range of settings, including military and civilian intelligence and law enforcement situations. Weiss (2008) discusses some of the more general issues regarding uncertainty and its communication in the intelligence domain.

Deception can take many different forms. The two most general varieties, however, involve denial or hiding evidence which would be valuable to an adversary, or providing false and misleading evidence which it is hoped that the adversary will observe and believe.

The primary conceptual framework which we employ for considering deception is reasoning under uncertainty. In particular, Bayesian networks (Pearl, 1988) provide a powerful modern tool for such reasoning, based on probability theory. We also discuss the Analysis of Competing Hypotheses (ACH) methodology (Heuer, 1999) which is well known to many US intelligence analysts. Since deception frequently involves the use of misleading evidence, we also briefly consider some characteristics of evidence and the combination of evidence.

2. BAYESIAN NETWORKS

Bayesian network models are powerful and flexible decision support tools, supporting a wide range of analyses. The framework is practically proven in diverse application areas such as medicine, forensic analysis and industrial fault diagnosis. BNs permit the fusion of disparate information, combining observations with subjective expert opinion. They are robust to missing information, facilitate value of information assessments and can represent variable source credibility, a key requirement for intelligence analysis.

At heart, a BN is a compact and efficient representation of a joint probability distribution over a domain of variables of interest. What makes it so powerful and flexible is the ease with which it supports different types of reasoning or inference. Furthermore, although we might expect probabilistic calculations performed over a domain of many variables to be slow and cumbersome, BN software typically employs some sophisticated algorithms, making use of local computations (Lauritzen and Spiegelhalter, 1988). These local computations, themselves made possible by conditional independence assumptions regarding the variables in the domain, avoid the need to work with the whole joint probability distribution when making inferences, thus speeding up the task considerably.

The qualitative structure of a BN is represented by a directed acyclic graph (DAG), portraying probabilistic dependencies and independencies within the domain. Each node in the graph represents a variable in the domain of interest. Although continuous variables are permitted, they are usually discretised so that each variable typically has a small number of mutually exclusive states which it can be in. An arc between two nodes indicates a direct probabilistic dependence between them, while the absence of an arc indicates a conditional independence relation. Hence, the DAG

ISBN 978-88-903724-3-8

contains a great deal of information, even before we consider any probability distributions. A fully specified BN, however, also requires the construction of conditional probability tables (CPTs) for each node. For parentless nodes, which have no arcs entering them, only a single prior marginal distribution has to be specified. For nodes with a single parent, a conditional probability distribution needs to be specified for each possible state of the parent variable. Finally, for chance nodes with several parents, a conditional probability distribution is usually required for every possible combination of parent states. While initially this may appear burdensome, in practice the requirement can often be relaxed, e.g. by making use of so-called Noisy-OR gates (Pearl, 1988) and their generalizations. This amounts to making certain reasonable independence assumptions, in exchange for a much simpler parameterization of the model.

There are many potential orderings of variables in a network, and the ordering chosen for a BN should represent the assumed dependencies and independencies as efficiently as possible. This usually means that the direction of an arc should follow the direction of causality when the relationship between two variables is causal. So, for example, it is the activities (or intent indicators) undertaken by a combat force which cause reports to be generated, the reports do not cause the activities to take place. Not all relationships in a BN have to be causal - weaker probabilistic dependencies will often be present. Exactly how such relationships should be represented and which way the arcs should be directed usually becomes clearer once the modeller has thought through their dependency implications. An invaluable guide in this respect is the d-separation criterion. See Pearl (1988) or Jensen (2001) for more details of this and for an introduction to Bayesian networks, more generally.

3. SCENARIO CONSIDERED

Here we employ a scenario described in McNaught et al. (2005). In it, a Blue HQ is trying to infer the intentions of a hostile Red force. The four possibilities considered are main attack (M), advance (A), defend (D) and withdraw (W). It is assumed that these are mutually exclusive, i.e. the Red force will only pursue a single course of action (CoA) at any given time. It is further assumed that each of these CoAs are equally likely at time zero, although this is not a general requirement and any prior distribution could be adopted. Several information cues or indicators of enemy intent are searched for by the Blue side in order to infer the Red CoA. Some cues may be associated with more than one CoA and some cues may be detected by more than one mechanism.

3.1. Bayesian Network of the Scenario

Space constraints prevent illustration of the entire BN so we present just a portion of it in Figure 1. This shows that the likely presence of indicators of enemy intent such as the establishment of airfields and counter-recce activities depend on the Red side's CoA. Furthermore, the probabilities of the two intent indicators displayed being observed or not by various mechanisms (e.g. air recce or ground recce) depend on the presence or absence of those indicators. Part of the timeline associated with this scenario is shown in Table 1.

Table 1: Timeline for Scenario.

Time Step	*Actions Taken by the Red Side and Indicators Detected by the Blue Side*
1	Blue establishes air and ground recce.
2	Red deploys air and ground recce as deception; Red increases counter-recce activities as deception; Red establishes dummy airfields as deception.
3	Red establishes demolition on bridges; Blue sub-unit reports sighting of Red recce (S3MA1); Blue ground recce reports Red counter-recce activities (S2MAD4); Blue air recce reports sighting of Red aux airfields (S1MA3).
4	Red conducts feint attacks; Blue ground recce report sighting of Red aux airfield (S2MA3) and demolition on bridges (S2DW15); Blue sub-unit reports local attacks (S3M8).
5	Red evacuates non-essential services; Blue sub-unit reports sighting of demolition on bridges (S3DW15).
6	Red employs smoke and jamming and a defensive frontage; Blue ground recce reports sighting of Red evacuation of non-essential services (S2W19) and Red's use of smoke (S2MW10); Blue sub-unit reports Red's use of smoke (S3MW10) and jamming (S3MW11); Blue Signals report Red's jamming (S4MW11); Blue sub-unit reports Red's defensive frontage (S3W18).
7	Red begins systematic destruction of bridges and commences withdrawal; Blue air and ground recce report sightings of Red destruction of bridges (S1W20 and S2W20).

3.2. Initial Results

As time progresses and fresh observations are made by the Blue side, the BN updates our belief in each Red CoA as shown in Figure 2.

The final probability distribution of 'Enemy Intent' is:

$$P(M \mid \text{All evidence}) = 0.39; P(A \mid \text{All evidence}) = 0.06;$$
$$P(D \mid \text{All evidence}) = 0.04; P(W \mid \text{All evidence}) = 0.51.$$

Although the BN eventually 'got it right', the true intent of the Red side only became apparent towards the end. For much of the time, 'Main Attack' seemed the likeliest CoA.

3.3. Results After Negative Evidence is Included

Events associated with one Red CoA or another which were not observed to take place were previously assumed unknown. Now, such events are treated as

ISBN 978-88-903724-3-8

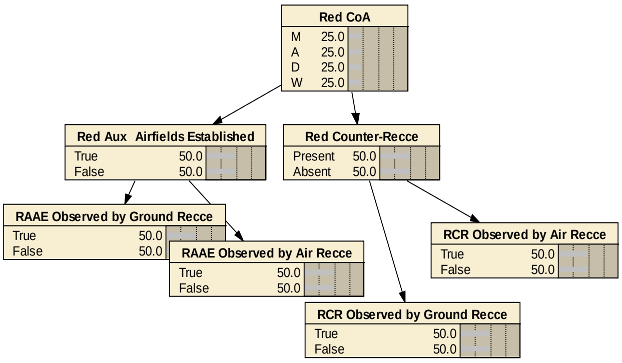

Figure 1: Partial BN of the scenario.

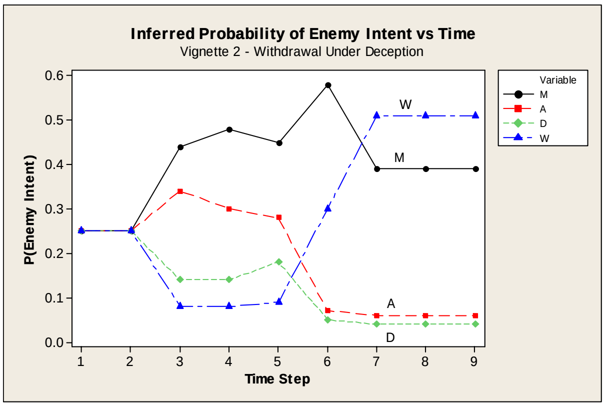

Figure 2: Probability Distribution of Enemy Intent vs Time

definitely not having occurred. The same underlying events are generated as in the first experiment, and the same positive intelligence reports are received at the same times. The difference is that in addition to the positive intelligence reports, there are now a number of 'negative' intelligence reports indicating that certain things have not been reported.

In deciding when to instantiate a report node with negative evidence, we have looked at the latest time we would expect a positive report to be received across the four possible states of Enemy Intent. If it has not been received by that time, we have instantiated a negative report for that indicator in the next time-step. The revised results for this scenario, incorporating the effects of negative evidence, are shown in Figure 3.

Clearly, this time the BN performs much better when the negative evidence is also taken into account. Firstly, the final distribution of 'Enemy Intent' is more decisive in each case. In Figure 3, the final distribution of enemy intent is now given by: $P(M \mid \text{All evidence}) = 0.01; P(A \mid \text{All evidence}) = 0.005;$ $P(D \mid \text{All evidence}) = 0.01; P(W \mid \text{All evidence}) = 0.975.$ Secondly, the correct option is identified earlier by the network. While it is difficult to quantify the benefit obtained by identifying the true enemy course of action sooner, this could be addressed in a simulation study.

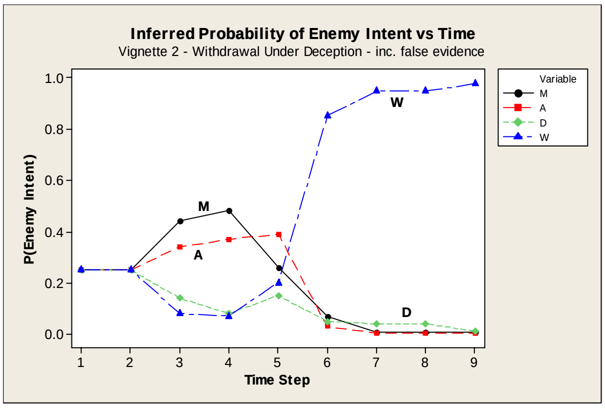

Figure 3: Probability Distribution of Enemy Intent vs Time with Negative Evidence Included

3.4. Conflict Measure as a Possible Indicator of Deception or a Missing Hypothesis

We investigate the use of a recognised BN conflict measure as a possible indicator of deception or a missing hypothesis. This measure is the ratio of the product of each piece of observed evidence's marginal probability to the joint probability of the observed evidence set, i.e.

$$\frac{P(e_1)P(e_2)...P(e_n)}{P(e_1, e_2, ..., e_n)}.$$

The rationale behind this ratio is that when the observed pieces of evidence are generally in agreement, i.e. taken together they form a coherent hypothesis, the evidence will tend to be positively correlated and so the joint probability of the various observations in the denominator will be greater than the product of the marginal probabilities in the numerator, leading to a ratio less than 1. A ratio in the region of 1 would only be expected if the evidence variables were largely independent of each other. However, a ratio greater than 1 implies that the joint probability of the observations is less than the probability of their occurrence if they were independent. In other words, the evidence does not paint a coherent picture, which might indicate in an adversarial context that a deception is being undertaken or that a more realistic hypothesis has not been considered.

Figure 4 shows how this ratio changes over time in this scenario both when only positive observations are taken into account and when negative evidence is also included. The ratio is notably lower when negative evidence is included.

Using the above example, we now consider the situation where the correct hypothesis is not being considered. Hence, the Red CoA 'Withdraw' is removed as a possible hypothesis. After making this change to the BN and then entering the same set of positive evidence as before, the final probability distribution over the remaining Red CoAs is:

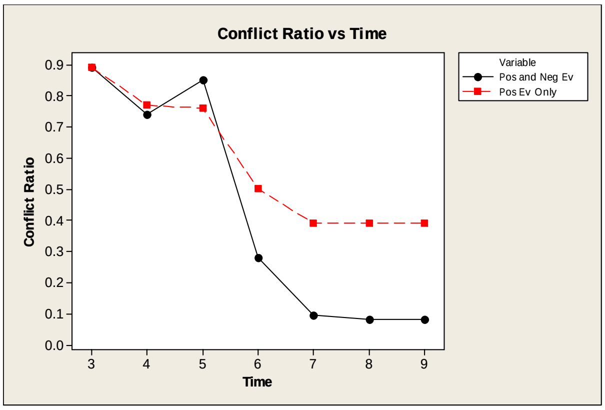

Figure 4: Conflict Ratio vs Time

P(M|Positive evidence) = 0.829; P(A|Positive evidence) = 0.103; P(D|Positive evidence) = 0.068. However, when negative evidence is also included, the final probability distribution becomes: P(M|All evidence) = 0.329; P(A|All evidence) = 0.209; P(D|All evidence) = 0.463. Clearly, when only positive evidence is considered, we may be mis-lead into believing that the Red CoA is a main attack. When negative evidence is included, however, the situation is much less clear with the single most likely hypothesis now being Red defence, the closest option to the true but unconsidered hypothesis of withdrawal. The conflict measure ratio is also plotted for these two cases in Figure 5. Note that this ratio never falls below 0.1. Although we cannot use such a threshold more widely, in a new situation we could possibly try to estimate it via simulation. A higher value could indicate, as here, that a new hypothesis needs to be considered, possibly one that is being masked by an adversary or otherwise seems implausible.

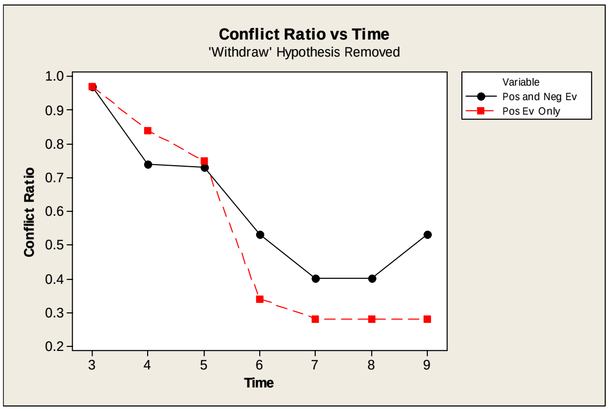

Figure 5: Conflict vs Ratio with the 'Withdraw' Hypothesis Removed

The next change that we consider is to introduce a new general hypothesis 'Other' while still leaving out the 'Withdraw' hypothesis. This represents a situation where the correct hypothesis is not among those being considered but nonetheless other possibilities are still being entertained. Such an approach is recommended within the ACH-CD method described in section 4.3. Results when only positive evidence is considered are presented in Figure 6. Results when negative evidence is included as well are presented in Figure 7.

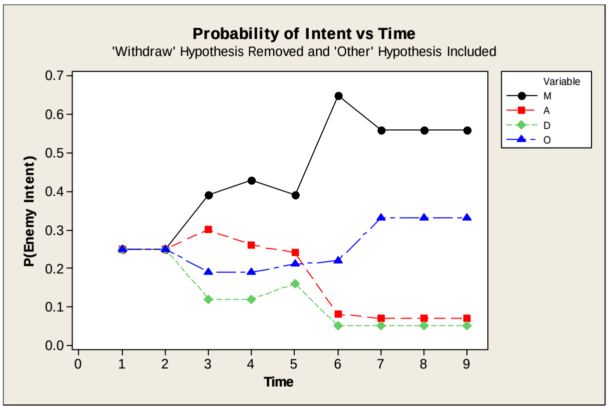

Figure 6: P(Enemy Intent) vs Time with Correct Hypothesis Removed and 'Other' Hypothesis Added

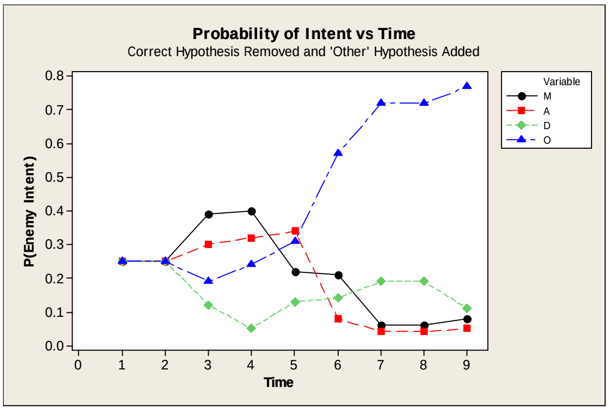

Figure 7: P(Enemy Intent) vs Time with Correct Hypothesis Removed and 'Other' Hypothesis Added

It is clear that the inclusion of negative evidence again improves the inference. The 'Other' hypothesis finishes strongly ahead of the remaining hypotheses, indicating that new possibilities need to be considered to explain the situation.

4. ANALYSIS OF COMPETING HYPOTHESES

Analysis of Competing Hypotheses (ACH) is a methodology developed by Heuer (1999) to help intelligence analysts overcome various cognitive biases, particularly confirmation bias. This is the tendency to overlook or underweigh evidence which contradicts the currently most favoured hypothesis, while overweighing supportive evidence. A matrix is developed in which the columns correspond to the set of plausible hypotheses and the rows correspond to items of evidence. The elements of the matrix record the extent to which each item of evidence supports or contradicts each hypothesis. Relevant negative evidence can and should also be included as rows in the matrix.

4.1. The ACH Framework

The basic outline of ACH is as follows:

1. Identify the alternative hypotheses to be considered.
2. Identify what evidence and assumptions are relevant to these hypotheses.
3. Construct the ACH matrix where the alternative hypotheses are the columns and each separate piece of evidence and assumption is a row.
4. In the matrix, indicate what evidence (including negative evidence) and assumptions supports or contradicts each of the alternative hypotheses, and by how much, removing that which does not discriminate between hypotheses.
5. Compare the relative likelihoods of all hypotheses, paying particular attention to evidence which contradicts a hypothesis, and identify future milestones when discriminating new evidence might come to light.

4.2. Partial ACH Example Matrix

Table 2 displays five example rows for an ACH matrix related to the scenario presented above. Evidence observations E1 and E2 are only weakly diagnostic as we might reasonably expect them to be present regardless of the Red CoA, although they are only slightly more likely for some CoAs than others. E3 is strongly associated with a Red Withdrawal. Failure to observe E4 (negative evidence) makes Defence or Withdrawal more likely than either offensive CoA while failure to observe E5 makes Defence a much less likely Red CoA. E6 is much less likely for the Withdrawal CoA but is employed as a means of deception in this scenario. While the Main Attack column may contain the most plus signs, note that overall, however, the Withdraw CoA column has the fewest minus signs and this is seen as more important in ACH. While human nature often makes us look for confirming positive signs, disconfirming negative signs may prove more valuable in many situations. This is a key motivation for the ACH framework and is particularly aimed at overcoming confirmation bias.

Table 2: Partial ACH Matrix

	Main Attack	**Advance**	**Defend**	**Withdraw**
E1: Red Radio Silence	++	+	++	+
E2: Red conducts feint attacks	++	+	+	+
E3: Red Evacuation of Various Services	-	-	-	++
¬E4:No observed forward movement of logistics	-	-	+	+
¬E5:No AT assets observed at frontline	+	+	--	+
E6: Red Counter-Recce Forces Operating	++	++	++	-

4.3. Combining ACH and BNs

ACH-CD (the CD standing for counter-deception) is an approach combining ACH and BNs, proposed by Elsaesser and Stech (2007). An example is provided concerning the Battle of Midway in which the position of US aircraft carriers is the basis of the deception. In another example concerning the D-Day landings, it is the transportable port facilities known as Mulberry which lies at the centre of the deception.

In their approach, a particular hypothesis of interest, H, is instantiated and the conditional probabilities of each piece of observable evidence given the hypothesis $P(e_i \mid H)$ recorded. Similarly, the condition ‘not H’ is then instantiated and the values of $P(e_i \mid \neg H)$ are obtained from the network. The ratio $\frac{P(e_i \mid H)}{P(e_i \mid \neg H)}$ indicates how important this piece of evidence is in discriminating between H and not H. In statistics, this ratio is well known as the likelihood ratio. Since it does not depend on the prior probability of the hypothesis, it is a direct measure of the weight of the evidence. For this reason, it has also become increasingly popular among forensic scientists, e.g. see Taroni et al. (2006).

5. EVIDENCE CHARACTERISTICS

A piece of evidence has many characteristics. These include relevance to the question being addressed, timeliness, since we might expect a newer observation to carry greater weight than an older one of the same type, and source credibility, particularly where human intelligence is involved. In reasoning about some adversary’s intentions, we frequently need to combine multiple pieces of uncertain evidence with different degrees of relevance, different time stamps and coming from sources with varying degrees of credibility. Identifying unreliable sources is particularly important to reduce vulnerability to deception, as is identifying common or highly dependent sources.

Deception may affect the evidence marshalling process, described by Schum (2001). This concerns the organisation of evidence to make a case and might include analysis of evidence gaps, and notions of evidence thresholds to take different actions such as more intrusive surveillance or making an arrest.

These aspects could again be modelled utilizing the framework of a static BN. In the experimental, visual analytic ‘Jigsaw’ system (Stasko et al. 2008) developed to help intelligence analysts navigate a vast array of potentially relevant documents, provision is made for a 'shoebox' which is essentially an evidence marshalling tool. Such a tool can help an analyst to organise the available evidence, so aiding the construction of a coherent case.

As well as organising available evidence, such a tool can help highlight gaps in the evidential support for a hypothesis. With often very limited resources, support is required to identify the most promising gaps to investigate. Probabilistic decision support tools such as Bayesian networks can help in such situations.

6. CONCLUSION

Deception fundamentally involves reasoning under uncertainty and probabilistic modelling therefore suggests itself as a potentially useful framework for modelling and trying to detect deception activities. In this paper, we presented a scenario involving a Red adversary with four possible courses of action available, representing the set of alternative hypotheses being considered. We showed how a Bayesian network of the situation could be constructed and used to update our belief distribution over the alternative hypotheses as new observations were made by our recce units and other assets. In particular, we noted that the incorporation of negative evidence alongside positive observations improved the ability of the BN to infer the correct course of action.

We also examined what would happen if the correct hypothesis was not even under consideration. When the correct 'Withdraw' hypothesis was simply removed, leaving the other three, the BN performed poorly with only positive evidence. It did improve, however, when negative evidence was also included. When a general alternative hypothesis of 'Other' was added to the three remaining CoAs, however, the BN performed better. Although still coming to the wrong final conclusion of 'Main Attack' when using only positive evidence, the 'Other' hypothesis was considered the second most likely with a sizeable posterior probability of 0.33. When negative evidence was also included, however, the hypothesis 'Other' ended up the most likely with a posterior probability of 0.77, leaving all other hypotheses trailing well behind.

While deeper and more wide-ranging investigation is required, this suggests that in highly uncertain situations, particularly where deception is prevalent, we should more routinely consider adding such an alternative to the set of hypotheses being considered.

REFERENCES

Elsaesser, C. and Stech, F., 2007. Detecting deception. In: Kott, A. and McEneaney, W.M., eds. *Adversarial Reasoning – Computational Approaches to Reading the Opponent's Mind*. Boca Raton, FL: Chapman & Hall, 101-124.

Pearl, J., 1988. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Mateo, CA: Morgan Kaufmann.

Heuer, R., 1999. *The Psychology of Intelligence Analysis*. Washington DC: Center for the Study of Intelligence, CIA.

Jensen, F.V., 2001. *Bayesian Networks and Decision Graphs*. New York: Springer. Mercer, P.A. and Smith, G., 1993.

Lauritzen, S.L. and Spiegelhalter, D.J., 1988. Local computations with probabilities in graphical structures and their application to expert systems. *Journal of the Royal Statistical Society, Series B*, 50, 157-224.

McNaught, K.R., Ng, B. and Sastry, V.V.S., 2005. Investigating the use of Bayesian networks to provide decision support to military intelligence analysts. In: Merkuryev, Y., Zobel, R. and Kerckhoffs, E.J.H. (Eds.), *Proc. 19th European Conference on Modelling and Simulation*, 72-79. June 2005, Riga, Latvia.

Schum, D.A., 2001. Evidence marshaling for imaginative fact investigation. *Artificial Intelligence and Law*, 9, 165-188.

Stasko, J., Gorg, C. and Liu, Z., 2008. Jigsaw: supporting investigative analysis through interactive visualization. *Information Visualization*, 7, 118-132.

Taroni, F., Aitken, C., Garbolino, P. and Biedermann, A., 2006. *Bayesian Networks and Probabilistic Inference in Forensic Science*. Chichester: Wiley.

Weiss, C., 2008. Communicating uncertainty in intelligence and other professions. *International Journal of Intelligence and Counter-Intelligence*, 21, 57-85.

AUTHOR'S BIOGRAPHY

Ken R. McNaught is a senior lecturer in Operational Research (O.R.) at Cranfield University's School of Defence and Security situated at the UK's Defence Academy in Shrivenham. He has an MSc in O.R. from Strathclyde University and a PhD in O.R. from Cranfield University. He leads the Decision Analysis and Risk Modelling Lab where his research interests include simulation, combat modelling and decision support, particularly making use of probabilistic graphical approaches such as Bayesian networks and influence diagrams. He also teaches on a number of specialized MSc courses, including Military Operational Research and Defence Simulation and Modelling.