COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science (2018) 1–24

ISSN: 1551-6709 online
DOI: 10.1111/cogs.12636

# Inference in the Wild: A Framework for Human Situation Assessment and a Case Study of Air Combat

Ken McAnally,[a,b] Catherine Davey,[c] Daniel White,[d] Murray Stimson,[a] Steven Mascaro,[e] Kevin Korb[e]

[a]*Aerospace Division, Defence Science and Technology Group, Victoria, Australia*
[b]*Melbourne School of Psychological Sciences, University of Melbourne*
[c]*Advanced VTOL Technologies*
[d]*ScienceFX*
[e]*Bayesian Intelligence Clayton, Victoria*

Received 4 June 2017; received in revised form 8 April 2018; accepted 24 May 2018

**Abstract**

Situation awareness is a key construct in human factors and arises from a process of situation assessment (SA). SA comprises the perception of information, its integration with existing knowledge, the search for new information, and the prediction of the future state of the world, including the consequences of planned actions. Causal models implemented as Bayesian networks (BNs) are attractive for modeling all of these processes within a single, unified framework. We elicited declarative knowledge from two Royal Australian Air Force (RAAF) fighter pilots about the information sources used in the identification (ID) of airborne entities and the causal relationships between these sources. This knowledge was represented in a BN (the declarative model) that was evaluated against the performance of 19 RAAF fighter pilots in a low-fidelity simulation. Pilot behavior was well predicted by a simple associative model (the behavioral model) with only three attributes of ID. Search for information by pilots was largely compensatory and was near-optimal with respect to the behavioral model. The average revision of beliefs in response to evidence was close to Bayesian, but there was substantial variability. Together, these results demonstrate the value of BNs for modeling human SA.

*Keywords:* Situation awareness; Decision making; Expertise; Mental models; Cognitive modeling

Correspondence should be sent to Ken McAnally, Melbourne School of Psychological Sciences, University of Melbourne, Parkville, Vic. 3010, Australia. E-mail: mcanally@unimelb.edu.au

## 1. Introduction

### 1.1. Situation assessment

Situation awareness is a key construct in human factors. A widely adopted definition of situation awareness is "the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning, and the projection of their status in the near future" (Endsley, 1995a). While previous research (Adams, 1998; Durso et al., 1995; Endsley, 1995b; Taylor, 1990) has focused on the assessment of situation awareness as a state of knowledge, the design of socio-technical systems and training to support good situation awareness would benefit from a better understanding of the cognitive processes by which situation awareness is generated (i.e., situation assessment [SA]), and how these processes interact with general knowledge, including causal knowledge and doctrine, and decision making.

SA is primarily a process of inferring the state of the environment from the available evidence. In the "fog of war," this evidence may be partial, uncertain, and/or deliberately misleading. SA comprises "extracting information from the environment, integrating this information with relevant knowledge to create a mental picture of the current situation, using this picture to direct further perceptual exploration in a continual perceptual cycle, and anticipating future events" (Directorate of Defence Aviation and Air Force Safety, 2010). An integrated framework for situation and decision assessment would also include the planning of interventions and the prediction of their effects (Fig. 1; adapted from the perceptual cycle of Neisser, 1976). The observe-orient-decide-act (OODA) loop (Boyd, 1995) commonly adopted by the combat aviation community is an alternate description, although it conflates the search and intervention cycles. Both the OODA loop and this general framework are, however, underspecified with respect to the nature of relevant

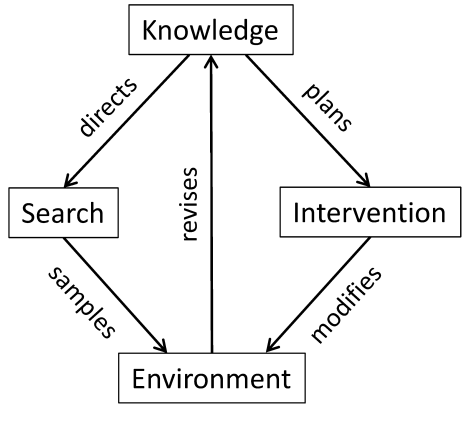


Fig. 1. Process model of situation and decision assessment (adapted from the perceptual cycle of Neisser, 1976).

knowledge structures, how search for information is directed, how knowledge is revised in the light of new information, and how interventions are planned.

### 1.2. Inference using causal models

People naturally make sense of the world by constructing causal mental models of generative processes within it (Gopnik et al., 2004; Sloman, 2005). Causal models support both predictive and diagnostic inference. Predictive inference is performed by propagating from observed causes to expected effects. Diagnostic inference is performed from observed effects back to potential causes. Causal models can guide the search for new information in the environment by specifying the value of different information sources. Causal models also support decision making whereby proposed interventions are conducted in a (mental or computer) simulation of the model and their effects are predictively inferred (Hagmayer & Sloman, 2009; Korb, Hope, Nicholson, & Axnick, 2004; Sloman, 2005; Sloman & Hagmayer, 2006). This is consistent with the mental simulation of possible actions by experienced operators in unfamiliar situations during naturalistic decision making (Klein, 1998).

Causal models may be implemented as Bayesian networks (BNs) where the strengths of the causal relationships are specified as conditional probabilities and the revision of beliefs occurs according to Bayes' rule (Bayes, 1763). BNs perform optimal inference under conditions of uncertainty (Pearl, 2009) and, therefore, provide a computational level of description (see Marr, 1982) of the inference task to be performed by humans because they define the normative mappings of information. Given the reported fallibility of human inference (e.g., Tversky & Kahneman, 1974), BNs have been proposed for use in computerized decision aids for SA in military command and control (Bladon, Hall, & Wright, 2002; Das, Grey, & Gonsalves, 2002).

### 1.3. Situation and decision assessment with BNs

BNs are attractive as "working models" that support all of the processes of SA (i.e., observation, revision of knowledge, directed search, and the planning of interventions) within a single, unified framework. In a BN, variables relevant to the domain are represented as nodes (e.g., Fig. 2a, boxes). Each node has a number of possible states that are meaningful in the context of the inference. Causal relationships between variables are represented by arrows (arcs) and associated conditional probabilities. The structure of the model, including the causal relationships, reflects *general knowledge* about the domain that may be instantiated in a mental model. The distributions of probabilities across the states of the variables (e.g., Fig. 2a, bars) represent *beliefs about the current situation* (i.e., situation awareness). As a scenario evolves and new information becomes available, beliefs throughout the model are revised according to Bayes' rule (Bayes, 1763) to accommodate both new and preexisting evidence.

In this simple example (adapted from Neapolitan, 1990), general knowledge about car mechanics is represented by the model shown in Fig. 2a. This model is admittedly

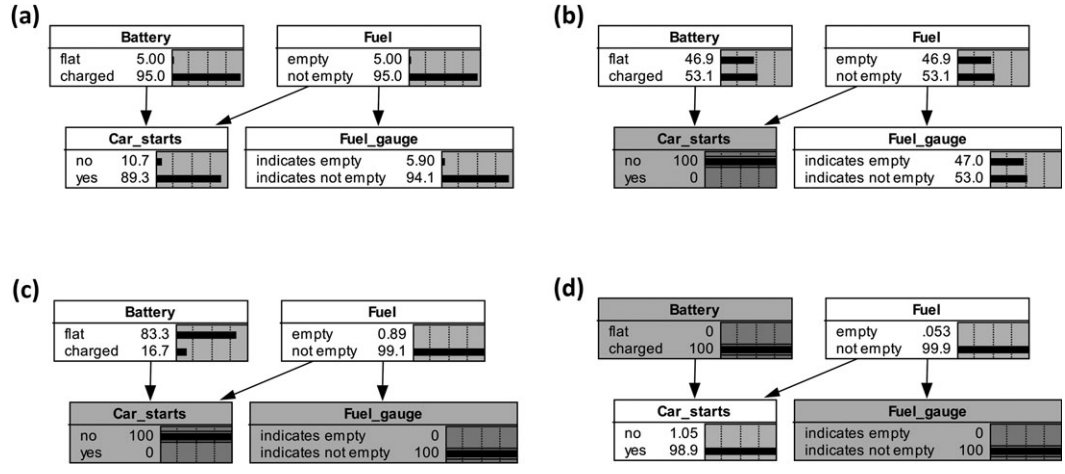


Fig. 2. SA using an example BN. (a) Simple causal model representing general knowledge of car mechanics. (b) Evidence that the car does not start revises upward the beliefs of an empty fuel tank and a flat battery. (c) Search for further information is directed by the model. The fuel gauge indicates not empty, revising downward the belief of an empty tank and further upward the belief in a flat battery. (d) A planned intervention to charge the battery is inferred to increase the probability of the car starting.

oversimplified in comparison to the mental models of most people, but it is sufficient to demonstrate how such models may be used. The distribution of probabilities represents the prior expectations before any evidence specific to a particular situation is attained (e.g., batteries are charged and fuel tanks are not empty most of the time). An observation that our car does not start is entered into the model (Fig. 2b) and is integrated with existing knowledge to increase the beliefs of both a flat battery (from 5.0% to 46.9%) and an empty fuel tank (from 5.0% to 46.9%; compare Fig. 2a and b). To diagnose the problem, we need to search for further information. Suppose we inspect the fuel gauge that indicates there is fuel (Fig. 2c). This new evidence is also integrated with current knowledge to revise downward the belief of an empty tank (from 46.9% to 0.89%) and revise upward the belief of a flat battery (from 46.9% to 83.3%; compare Fig. 2c and b). Interventions are planned by simulating a counterfactual world in which the intervention is conducted and observing the consequences. In this example, a proposed intervention to charge the battery (Fig. 2d) is expected to increase the probability that the car will start.

The present study examined whether SA by experienced military pilots could be modeled by BNs. In particular, the following questions were addressed: Are the knowledge structures supporting SA causal? How does knowledge direct information search? How is new information integrated with existing knowledge to revise beliefs? Although not examined here, this framework may also be applied to examine how interventions are planned (cf., Fig. 1).

This study examines what is perhaps the most fundamental of inferences made by military personnel; that is, the identification (ID) of entities as either friendly, neutral, or hostile. First, we conducted a task analysis of airborne threat ID by fighter pilots in a

defensive counter-air scenario. This analysis included the relevant sources of information, including context and data from cockpit systems. From this task analysis, we generated a causal BN model of the ID task. We then validated this model by conducting a low-fidelity simulation that required fighter pilots to perform ID (without the support of other entities such as Airborne Early Warning and Control). Search for information by the pilots and their evolving inference of ID are first presented descriptively and then modeled by a second, better-fitting BN model.

## 2. Methods

### 2.1. Domain elicitation from subject-matter experts

Two experienced F/A-18 pilots from the Royal Australian Air Force (RAAF) participated in structured interviews to elicit the variables relevant to the task of identifying airborne entities, including context variables and evidence available from cockpit systems. The possible states of each variable that were meaningful to the pilots in the context of the task and the presence of causal relationships between these variables were also elicited. The pilots then jointly estimated the probability of observing each state of a variable for each combination of states of any variables that were directly causal to it. The *declarative model* is defined by the elicited structure and conditional probabilities (Fig. 3). Preliminary validation of the declarative model was conducted in a series of informal scenario walk-throughs with the two pilots from whom the model was elicited and one other experienced F/A-18 pilot who had visibility of the model. These scenario walk-throughs were conducted by sequentially setting node states of hypothetical scenarios and evaluating the model's inference of ID for sensibility. Some minor adjustments were made to the parameters (i.e., conditional probabilities) of the model so that inferences of ID by the model were as expected.

### 2.2. Simulation experiment

We investigated whether the skilled behavior of pilots in a simulated cockpit environment (Fig. 4) was consistent with the declarative model elicited during interviews. Twenty scenarios were generated by probabilistic sampling from the declarative model. As this model represents the declarative knowledge of subject-matter experts, these scenarios are expected to approximate a sample of real-world circumstances. A limitation of such sampling is that only common scenarios are expected to be represented in a small sample. We, therefore, generated an additional 19 scenarios by experimenter selection of node states other than ID. One of these scenarios was selected to be highly ambiguous between hostile and neutral ID. The others were generated in pairs that differed by the state of a single node, where the change in that node resulted in a large change in inferred ID. These scenarios sometimes included rare and/or conflicting evidence.

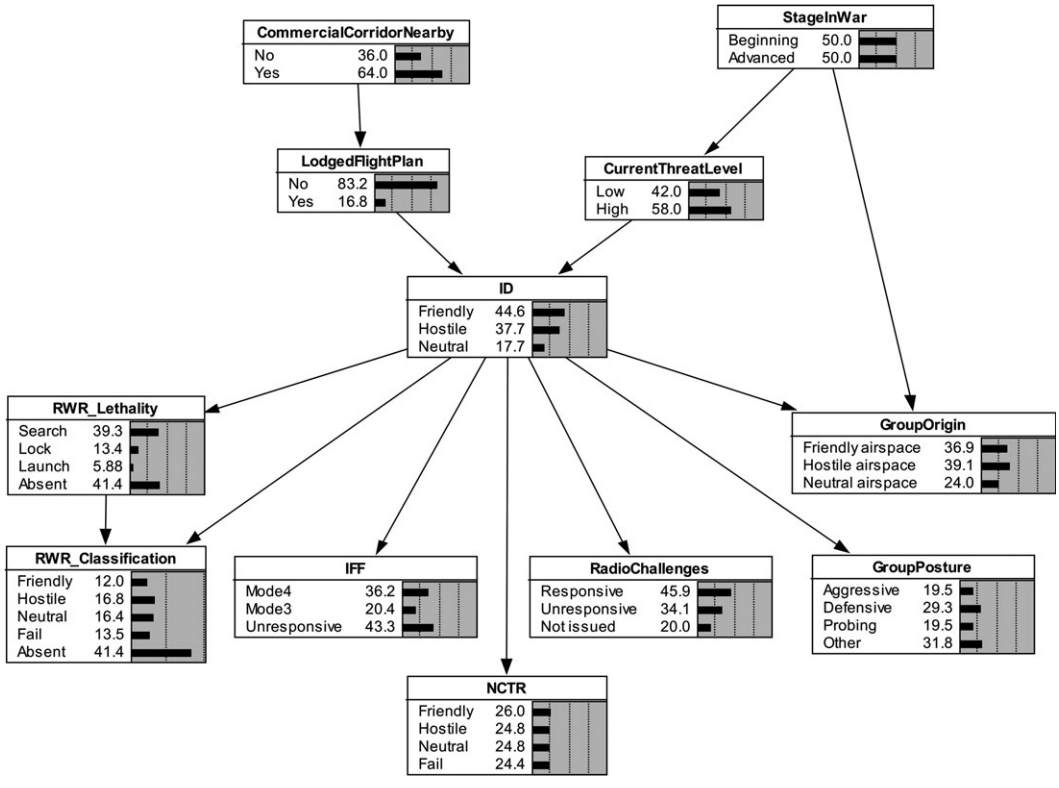


Fig. 3. Declarative model of entity identification. The structure of the model represents general causal knowledge about the domain, and the distribution of probabilities (bars) represents beliefs about the current situation. Context nodes (with arrows into the ID node) are causal of the entities airborne (ID node), which are, in turn, causal of evidence available to the pilot (with arrows from the ID node).

Except for ID, the state of each node identified in the declarative model was made available to the pilots on their request. This information was presented ecologically in the form of a simulated cockpit display. The pilots did not have access to the structure, conditional probabilities, or inferences of the model. The pilots' task was to identify unclassified airborne entities from context information available in a preflight brief and from evidence available from their cockpit systems. Cockpit-sourced information was only made available to the pilot upon request by a mouse click. Pilots were instructed to search each of seven available information sources in the order of perceived utility for inferring ID. In each step of a scenario, pilots selected a source of information, and then indicated their revised belief with regard to ID by moving a mouse cursor within a triangle in the top center of the display. Their degree of belief in a particular state of ID (friendly, hostile, or neutral [either civil or military]) was indicated by positioning the cursor in relative proximity to the vertices of the triangle. The probability distribution of the belief indicated by the cursor position was also presented numerically. This sequential

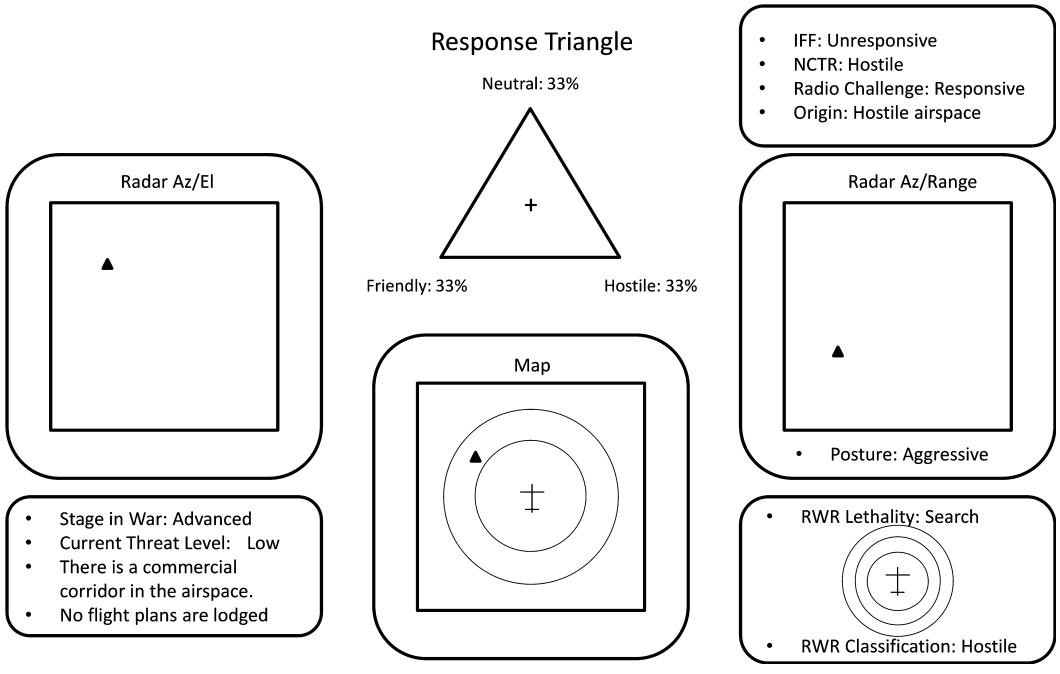


Fig. 4. Schematic of the low-fidelity simulation of F/A-18 cockpit. Context was given in the preflight brief (bottom left). Group posture was shown graphically on the radar displays (left and right), on the map page (bottom center), and in text in the radar azimuth/range display. RWR classification and RWR lethality were shown on the RWR display (bottom right). Other evidence was shown in text (upper right). Individual pieces of evidence were revealed sequentially in response to pilot request. Pilots indicated their current belief about target ID by moving the cursor in the response triangle (upper center).

search and response allowed the development of pilots' beliefs about ID to be tracked as evidence was accrued and integrated throughout a scenario.

Nine current F/A-18 pilots from RAAF Squadrons 3 and 77 (with from 300 to 2,400 h on type) completed 20 scenarios generated probabilistically from the declarative model. Another 10 current F/A-18 pilots from RAAF Squadrons 3 and 77 (with from 200 to 2,500 h on type) completed another 19 scenarios. In total, there were 2,960 judgments of target identity, where each judgment required pilots to state the degrees to which they believed the target to be friendly, hostile, and neutral.

All pilots gave informed consent. The experiment was given ethics approval by the Chief of Air Operations Division, Defence Science and Technology Organisation, in accordance with national guidelines.

### *2.3. Model selection*

The goodness of fit of the declarative model to the pilot responses was assessed by calculating the Kullback–Leibler (KL) divergence of the model predictions from the pilot

responses. KL divergence is equivalent to the average negative log likelihood (Shlens, 2014) and varies from zero for perfect agreement to infinity. Given that people often have difficulty in estimating probabilities (Tversky & Kahneman, 1974), we also assessed a model that retained the structural form of the declarative model, but estimated the conditional probabilities from the pilot responses using expectation maximization. The ability of this model to predict pilot responses was assessed by calculating the average KL divergence in a five-fold cross-validation ($KL_x$). Each model was trained on 80% of the data and tested on the remaining 20%. This was done five times with rotation of the test data such that each datum was represented in one test set. Since the eight successive search steps within each trial (i.e., one pilot conducting one scenario) were highly correlated, all data from a trial were kept together and included in either the training or the test set. In total, there were 370 trials.

We also conducted an exhaustive search across all simpler models where each node (except ID) was either present or absent. If present, it was assigned to the same causal tier (i.e., context or evidence) as in the declarative model. We forced a link between RWR classification and RWR lethality when these nodes were both present to capture the constraint that the "absent" state was the same for these nodes. We also allowed for the inclusion of an additional node (Pilot) as a cause of ID to capture individual variation across pilots. In total, 4,096 models were compared. The ability of each of these models to predict pilot responses was assessed by calculating the average KL divergence in a five-fold cross-validation. The model that best predicted pilot responses is henceforth referred to as the *behavioral model*. All BNs were programmed using Netica (Norsys) software.

## 3. Results

### 3.1. The declarative model

The declarative model is shown in Fig. 3. Context variables (commercial corridor nearby, lodged flight plan, stage in war, current threat level) have a causal influence on the type of aircraft likely to be in the air. The actual ID of the target results in evidence that is available from cockpit systems (e.g., identification friend or foe [IFF], radar warning receiver [RWR], noncooperative target recognition [NCTR]) and from the behavior of the entity (radio challenges, group posture, group origin). Group origin is the location of the group when first contact was made—not the location from which they took off.

The utility of each of these nodes for inferring ID is given by its mutual information (MI) with ID (Table 1). MI is the expected reduction in the entropy of ID resulting from evidence for that node and is dependent on the state of evidence for other nodes. The values in Table 1 represent MI with ID before any context or evidence is accrued.

It should be noted that as the declarative model was elicited from two subject-matter experts, it is expected to approximate real-world circumstances, but it may not represent them with high fidelity. Extensive engineering data and/or real-world observations would

Table 1
Mutual information of each node in the declarative model with ID (in the absence of evidence)

| Information Source | MI with ID (%) |
|---|---|
| IFF | 46.7 |
| Group origin | 18.5 |
| Group posture | 14.7 |
| Radio challenges | 14.1 |
| RWR classification | 10.3 |
| RWR lethality | 8.4 |
| Current threat level | 7.2 |
| NCTR | 1.5 |
| Stage in war | 1.4 |
| Lodged flight plan | 1.0 |
| Commercial corridor nearby | 0.01 |

be required to generate a model that is normative with respect to real-world circumstances. The declarative model was used to generate evidence in 20 of the scenarios for the simulation experiment.

### 3.2. Pilot responses in simulation

The design of the simulation experiment allowed pilots to revise their beliefs about ID as additional evidence became available throughout each trial. An example belief trajectory for a scenario is shown in Fig. 5. At step 0 (following the contextual information only), the pilot did not revise his belief from the uniform prior (i.e., equal probabilities of friendly, hostile, and neutral). As evidence became available during his search, the pilot revised his belief to be more certain of neutral. He did not revise his belief in the last two steps.

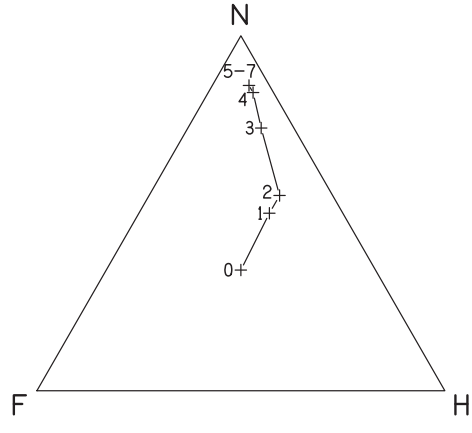


Fig. 5. Example of a pilot's belief trajectory during the course of scenario 13. F = friendly, H = hostile, and N = neutral. The search steps are numbered with 0 being the belief from the context only.

On 53% of trials, pilots did not revise their belief from the uniform prior probabilities in response to context. There was variability in the order in which pilots searched for further information, but a strong trend was observed for group posture, IFF, and group origin to be searched first, second, and third, respectively, and for NCTR to be searched last (Table 2).

Pilot search was largely compensatory in that participants generally revised their beliefs in response to each piece of evidence. When data were pooled across pilots, scenarios, and search steps, pilot beliefs were revised on 76% of occasions. Each revision of belief can be specified as a distance moved in the response space. The KL divergence is a measure of this distance and reflects the information about perceived ID gained during each revision. The average revision of beliefs was relatively small for step 0 following context information, relatively larger for steps 1 and 2, and then smaller again later in the search (Fig. 6). The average KL divergence of pilot beliefs differed significantly across steps, $F(3.0, 54.0) = 6.26$, $p = .001$, $\eta_p^2 = 0.26$.

On a small number of trials, pilot search was noncompensatory. Nine pilots reached certainty in their beliefs during the course of some (on average 2.4) scenarios and did not revise their beliefs further. In the most frugal of these searches, ID was classified friendly if IFF was Mode 4. According to the declarative model, there is a probability of 0.99 of friendly ID if IFF is Mode 4 in the absence of other information. Noncompensatory search was also occasionally observed for hostile identifications. In the most frugal of these, ID was classified as hostile if IFF was unresponsive, group origin was hostile airspace, and group posture was aggressive. According to the declarative model, these states result in a probability of .96 of hostile ID in the absence of other information.

The final belief of each of the nine pilots at the end of the 20 scenarios generated probabilistically from the declarative model is shown in Fig. 7 (crosses). Pilot beliefs were highly consistent for some scenarios (e.g., scenarios 15 and 20), but they differed greatly for others, despite a uniform training syllabus and access to the same context and evidence. Variation in pilot belief was mainly along the friendly neutral or hostile-neutral axis. It should be noted that there was no "true" ID in these scenarios. Rather, the

Table 2
Proportion of scenarios pooled across pilots in which each information source was inspected in each step

| Source | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| Group posture | **0.82** | 0.12 | 0.01 | 0.05 | 0 | 0 | 0 |
| IFF | 0.12 | **0.62** | 0.19 | 0.04 | 0.03 | 0 | 0 |
| Group origin | 0.06 | 0.19 | **0.48** | 0.10 | 0.15 | 0.02 | 0.01 |
| Radio challenge | 0 | 0.05 | 0.05 | **0.28** | 0.16 | **0.27** | 0.19 |
| RWR lethality | 0 | 0.02 | 0.07 | **0.29** | 0.18 | **0.29** | 0.15 |
| RWR classification | 0 | 0 | 0.05 | 0.13 | **0.38** | 0.21 | 0.23 |
| NCTR | 0 | 0 | 0.16 | 0.13 | 0.09 | 0.20 | **0.41** |

*Note.* The largest values for each source and step are shown in bold. Where two values are similarly large, both are shown in bold.

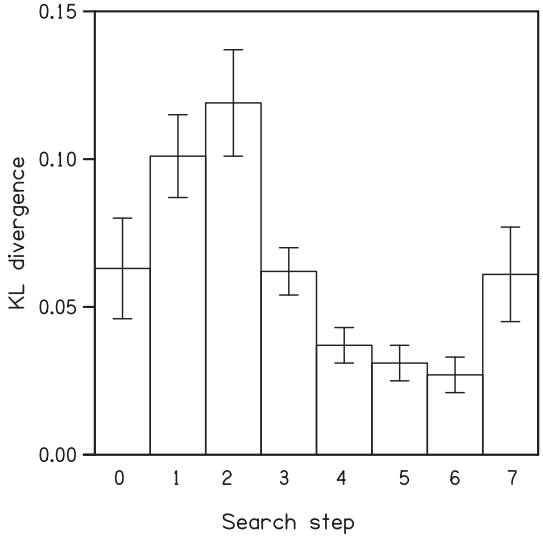


Fig. 6. Histogram of the KL divergence of pilot beliefs at each search step. Error bars represent standard errors of the mean, corrected for the repeated-measures design (Cousineau, 2005).

available evidence was consistent with friendly, hostile, and neutral ID to differing extents as indicated by the declarative model (see below).

Of particular interest for understanding friendly fire incidents, there were two scenarios (36 and 38) where there was substantial variation in pilot response across the friendly hostile axis (Fig. 8). The node states for these scenarios are shown in Table 3. In scenario 36, one pilot responded 0.81 probability of hostile ID and another responded 0.83 probability of friendly ID. The pilot responding hostile made an early judgment of 0.84 probability of hostile due to an unresponsive IFF and did not revise his beliefs further. The pilot responding friendly initially responded 0.86 probability of friendly due to defensive group posture. This was revised downward to 0.57 probability of friendly upon receipt of an unresponsive IFF, but subsequently revised upward again. Similarly, in scenario 38, one pilot responded with 0.90 probability of hostile ID and another responded a 0.90 probability of friendly ID. The pilot responding hostile made an early judgment of 0.89 probability of hostile ID when presented with an unresponsive IFF and did not revise that belief when presented with further information. The pilot responding friendly initially responded around .80 probability of hostile ID until the last step where group origin was found to be friendly airspace, resulting in a revision of belief to 0.90 probability of friendly ID. For each scenario, the context and evidence were consistent with a high probability of friendly, as indicated by the declarative model (Fig. 8; open squares).

### *3.3. Modeling SA*

We investigated whether SA by pilots could be well described by a BN. In doing so, we acknowledge that revisions of belief by pilots do not necessarily conform to Bayes'

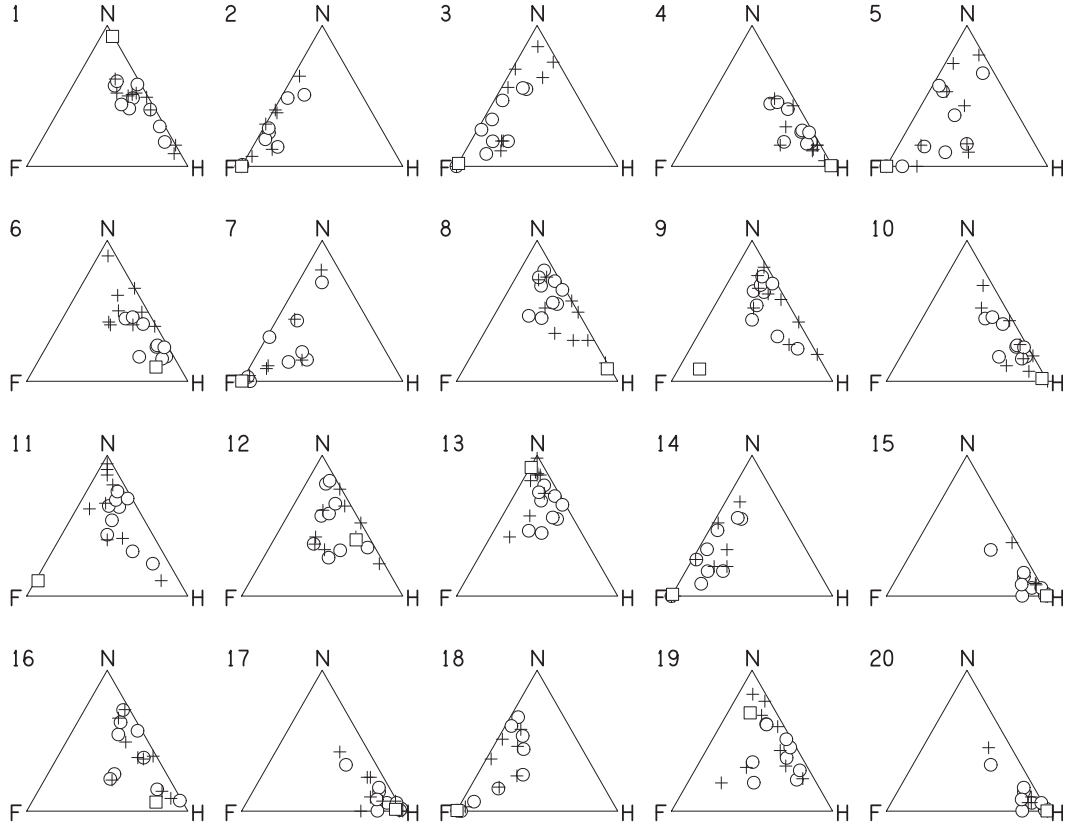


Fig. 7. Final beliefs of nine pilots (crosses), the declarative model (open squares), and the behavioral model (open circles) for the 20 scenarios generated probabilistically from the declarative model. F = friendly, H = hostile, and N = neutral.

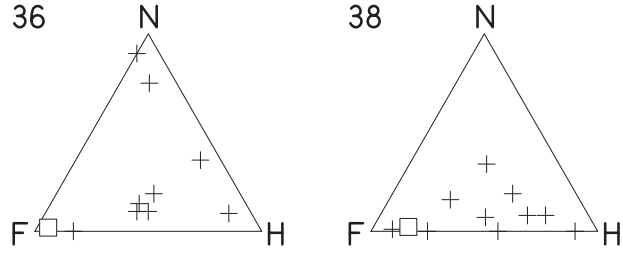


Fig. 8. Final beliefs of 10 pilots (crosses) and the declarative model (open squares) for scenarios 36 and 38. F = friendly, H = hostile, and N = neutral.

rule. Modeling of pilot inference with a BN has a number of advantages. First, the 2,960 observed judgments of ID (each with 2 degrees of freedom) can be reduced to a smaller number of model parameters. Second, the ability of pilots to discriminate between their

Table 3
Information states for scenarios 36 and 38

| Information Source | Scenario 36 | Scenario 38 |
|---|---|---|
| Stage in war | Advanced | Advanced |
| Current threat level | Low | Low |
| Commercial corridor nearby | No | No |
| Lodged flight plan | No | No |
| Posture | Defensive | Aggressive |
| IFF | Unresponsive | Unresponsive |
| Group origin | Friendly airspace | Friendly airspace |
| RWR lethality | Absent | Absent |
| RWR classification | Absent | Absent |
| Radio challenge | Not issued | Unresponsive |
| NCTR | Neutral | Friendly |

cognitive categories of friendly, hostile, and neutral entities may be estimated from that of the model. Third, information search, pilot beliefs, and the revision of those beliefs may be evaluated against those of the model. It should be noted, however, that such a BN will only perform the *correct* inference to the extent that pilots make the correct inference.

We recorded the search choices of each pilot and tracked his beliefs throughout each scenario (Fig. 5). For each combination of pilot and scenario, we provided the declarative model with the same order of evidence as that selected by the pilot and tracked the beliefs of the model. Fig. 7 shows the final beliefs of the declarative model from which the scenarios were generated (open squares). As described above, this model is expected to loosely approximate real-world circumstances because it is based on the judgments of only two subject-matter experts. On average, pilots responded similarly to the declarative model for about half of the scenarios, but for others there was a large divergence (e.g., scenarios 9 and 11). The ability of the declarative model to predict inference of ID by the pilots was evaluated by calculating the average KL divergence of the model responses from the pilot responses across pilots, scenarios, and search steps (2,960 observations). The declarative model had only a moderate ability to predict pilot responses in the simulation as indicated by an average KL divergence of 0.59.[1] Another measure of the agreement between model predictions and pilot responses is the intraclass correlation for absolute agreement. This correlation was 0.71, indicating a moderate agreement between ID probabilities predicted by the declarative model and the responses of the pilots in the simulation.

We evaluated another model that retained the structure of the declarative model, but where the conditional probabilities were estimated from pilot performance. To avoid overfitting the data, the ability of this model to predict pilot performance was assessed in a five-fold cross-validation. The average KL divergence of the re-parameterized model from the pilot responses was 0.20. This is substantially lower than for the declarative model, indicating that pilot behavior did not conform to the probabilities elicited for the declarative model.

We also tested models that differed in structure from the declarative model, where each node (except ID) could be either present or absent. If present, it was assigned to the same causal tier (i.e., context or evidence) as in the declarative model. An additional causal node (Pilot) was allowed to account for individual variation across pilots. An exhaustive search was conducted across all models meeting these constraints. Each of these models was evaluated by five-fold cross-validation. The *behavioral model* (Fig. 9) is that which had the lowest KL divergence of model predictions from the pilot responses (0.14). The intraclass correlation coefficient between model and observed probabilities of ID was 0.86, reflecting good absolute agreement between the model and the pilot beliefs. The final beliefs of the behavioral model are shown as open circles in Fig. 7 and show a similar distribution to the pilot beliefs.

The behavioral model is essentially a naive Bayes model (Mitchell, 1997) with three independent attributes of ID (IFF, group posture, group origin). The link between Pilot and ID allows for individual differences in the prior belief of ID (i.e., in the absence of evidence; Fig. 10). The relationship between ID and its three attributes is the same for all pilots.[2] The behavioral model has two free parameters unique to each pilot and another 23 that are common across pilots. The utility of each node in the behavioral model for inferring ID is given by its MI with ID (Table 4). These scores are considerably lower than the corresponding scores for the declarative model (Table 1).

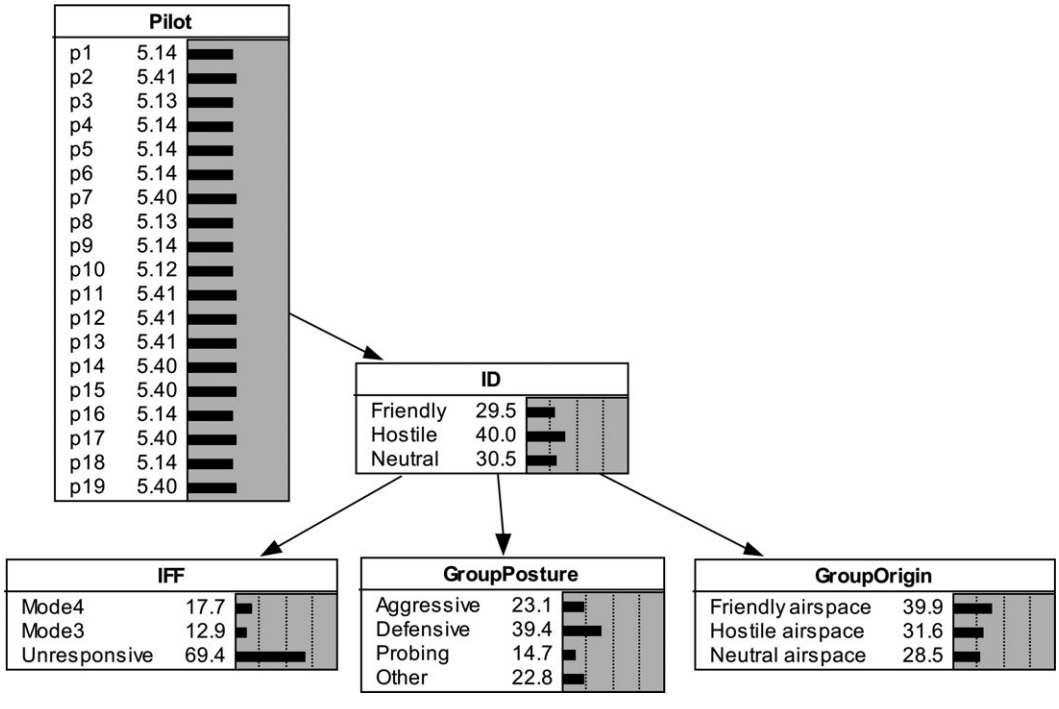


Fig. 9. The behavioral model.

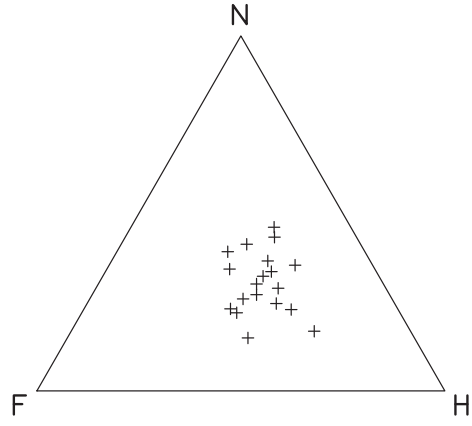


Fig. 10. Distribution of prior beliefs of ID across pilots in the behavioral model.

Until now, we have described and modeled the distribution of inferences of ID, conditional on the context and evidence. We can use the behavioral model to investigate a more general question: How well could pilots discriminate between their cognitive categories of friendly, hostile, and neutral across the range of possible scenarios? We simulated 100,000 friendly, hostile, and neutral entities in the behavioral model to generate evidence (IFF, group origin, and group posture) and then examined the recovery of ID by the same model.[3] While the majority (i.e., 56.9% of friendly, 50.8% of neutral, and 70.5% of hostile cases) of inferred IDs were correct, there were a substantial number of incorrect inferences, including some with high confidence (Fig. 11; only the first 300 cases are plotted for clarity). Of particular interest, 0.14% of friendly and neutral entities were inferred with high confidence (>0.9) to be hostile and 1.6% of hostile entities were inferred with high confidence (>0.9) to be not hostile.

### *3.4. Information search*

Knowledge may be used to direct search for information by specifying the utilities of each of the search options. According to the behavioral model, the utility of any source in inferring ID is reflected by the MI of that source with the ID node. According to the behavioral model, IFF should be searched first. For some scenarios, group posture and group origin should be searched second and third, respectively. For other scenarios, the order of the latter two sources should be reversed. In general agreement with the model, there was a strong trend for group posture, IFF, and group origin to be searched first, second, and third, respectively (Table 2). In the simulation, group posture was indicated by the tactical picture in the radar displays and by a text label. While group posture was less informative about perceived ID than was IFF, it enabled pilots to orient themselves at the beginning of each scenario.

We investigated the relationship between an information source's MI with ID and the step in which it was selected during search. On average, early search steps interrogated

Table 4
Mutual information of nodes in the behavioral model with ID (in the absence of evidence)

| Information Source | MI with ID (%) |
|---|---|
| IFF | 7.8 |
| Group origin | 6.3 |
| Group posture | 6.0 |
| Pilot | 2.5 |

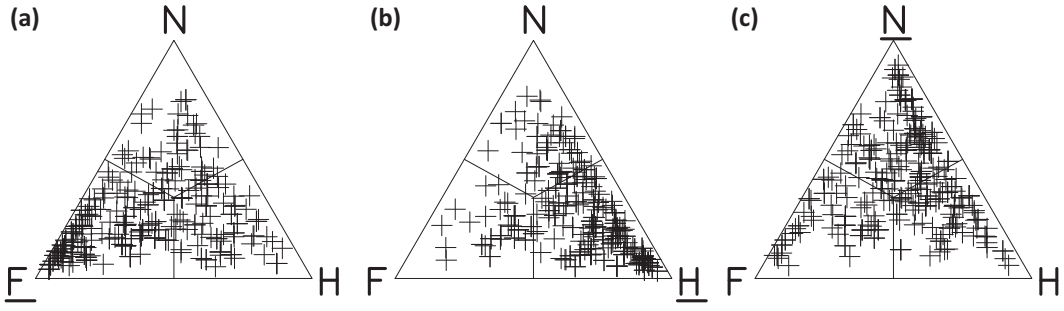


Fig. 11. Inference by the behavioral model where the simulated ID is (a) friendly, (b) hostile, or (c) neutral. For clarity, only 300 cases for each ID type are shown. The simulated ID is indicated by underlining the F, H, or N label. The lines within the triangles indicate classification boundaries.

sources with higher MI than later steps (Fig. 12: pilot choice), $F(2.6, 46.2) = 94.7$, $p < .001$, $\eta_p^2 = 0.84$. It should be noted that a random search order is also expected to result in reductions in MI toward the end of the search where ID is more constrained by the evidence already available, $F(4.0, 72.7) = 2.84$, $p = .03$, $\eta_p^2 = 0.14$ (Fig. 12: random choice). Crucially, there was a significant interaction between Choice and Step, $F(3.1, 55.4) = 57.4$, $p < .001$, $\eta_p^2 = 0.76$, indicating that pilots were sensitive to MI, or a correlate of MI, in their search for information.

The efficiency of information search is reflected in the rate at which uncertainty (entropy) about ID is reduced. We calculated the reduction in ID entropy at each search step, normalized to the total reduction in entropy of ID for each scenario (Fig. 13).[4] For comparison, we also plotted the reductions in ID entropy that would result from a random search and an optimal search. In optimal search, the remaining information source with the highest MI is chosen at each step. For all search steps, the average reduction in ID entropy for the pilot choice was higher than for a random choice, $t(18) \geq 2.66$, $p \leq .016$. For only the first three search steps was the average reduction in ID entropy for the pilot choice significantly lower than for the optimal choice, $t(18) \geq 2.16$, $p \leq .045$.

According to the behavioral model, search should be noncompensatory, that is, it should be terminated after the interrogation of IFF, group posture, and group origin with no subsequent revision of beliefs. As discussed above, there was a strong trend for these sources to be searched first, but pilots did revise their beliefs in the light of information from other sources (Fig. 6).

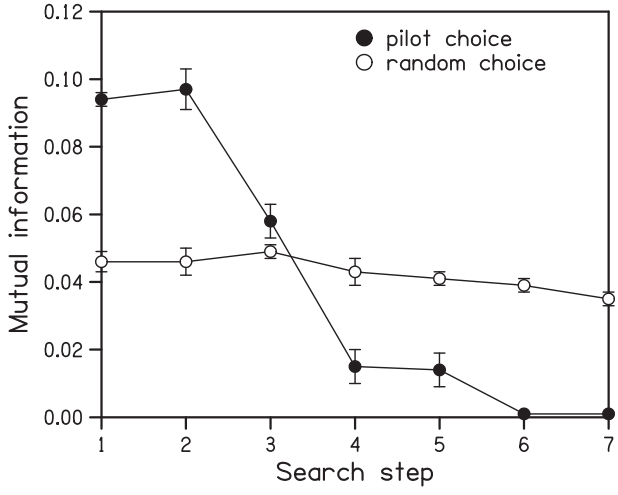


Fig. 12. Plot of the MI between the information source chosen at each search step and ID. Error bars represent standard errors of the mean, corrected for the repeated-measures design.

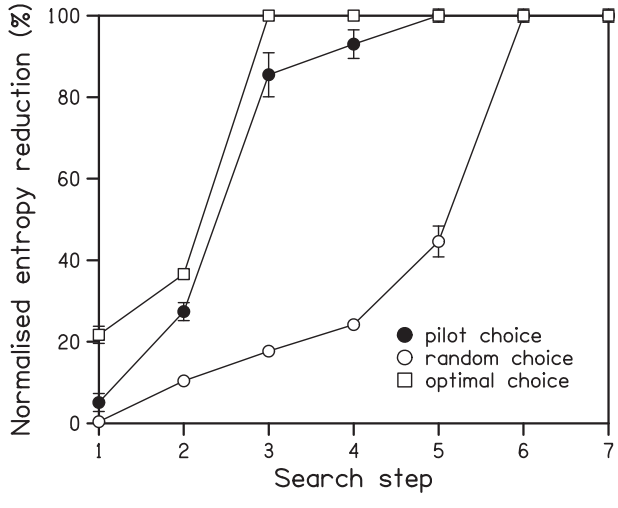


Fig. 13. Plot of cumulative reduction in ID entropy at each search step. Error bars represent standard errors of the mean, corrected for the repeated-measures design.

### *3.5. Revision of beliefs*

Bayes' rule (Bayes, 1763) specifies how beliefs should be revised in the light of evidence. This rule is expressed most simply if beliefs are expressed as odds, that is, the ratio of probabilities of two competing hypotheses.

$$\text{Revised odds} = \text{prior odds} \times \text{LR of evidence}$$

where LR is the likelihood ratio, that is, the ratio of probabilities of observing the evidence under each hypothesis. It is, therefore, possible to examine the beliefs (odds) separately from the manner in which they are revised (the LR).

Odds (friendly/hostile and neutral/hostile pooled) and LRs for pilot responses are plotted against those for the behavioral model in Fig. 14. Consistent with the good fit of the behavioral model to pilot beliefs discussed earlier, there was good absolute agreement between the model and the pilot belief odds (intraclass correlation = 0.79). In contrast, the absolute agreement between the model and the pilot LRs was lower (intraclass

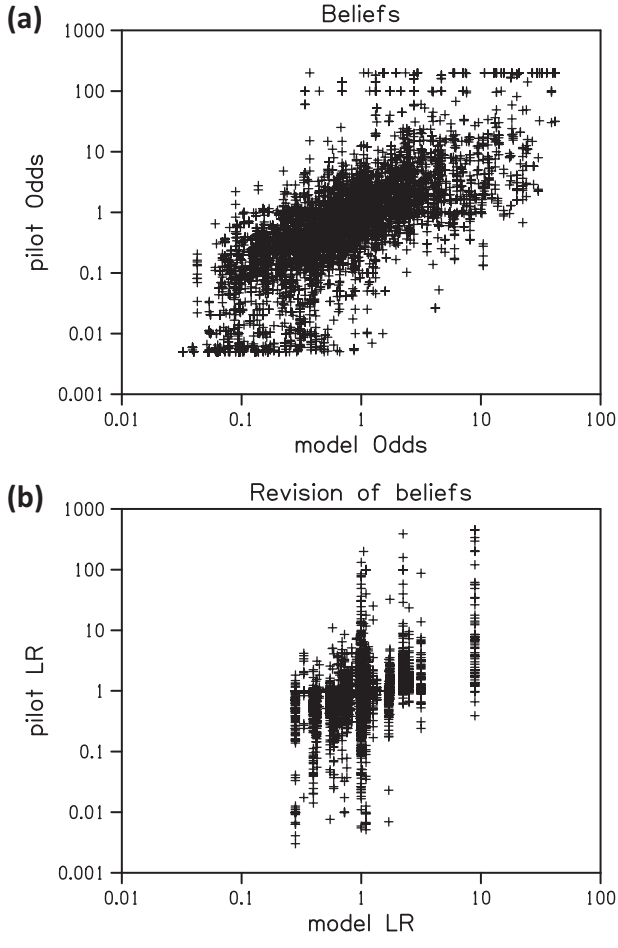


Fig. 14. Scatterplot of (a) pooled odds of F/H and N/H and (b) pooled likelihood ratios of F/H and N/H for pilot responses against behavioral model responses. Values greater than 1 favor friendly or neutral and values less than 1 favor hostile.

correlation = 0.51). Pilots demonstrated a large range of beliefs and large revisions of those beliefs, but the model was more moderate, particularly for revisions of belief. That the range of belief revision by the model was smaller than the range of beliefs reflects the consistent nature of the evidence within each trial, compared to that across trials. That the behavioral model fits the beliefs better than the revisions of beliefs is consistent with maximizing the likelihood of the model, given the data (i.e., beliefs) during the learning of the model conditional probabilities.

## 4. Discussion

It was possible to elicit from pilots a detailed causal model of inference about the ID of airborne entities. This declarative model had a moderate ability to predict the behavior of pilots during a low-fidelity simulation. However, behavioral data were required to develop more robust models of pilot behavior in simulation. As expected, the predictive ability of the declarative model was improved when conditional probabilities were estimated from pilot performance rather than from elicitation in interview. This is consistent with the hypothesis that people have difficulty in estimating probabilities upon which their behavior is based (Tversky & Kahneman, 1974).

Twenty of the scenarios were generated probabilistically from the declarative model in order to sample the expected most common scenarios. In theory, this sampling may have constrained the resulting behavioral model to share features with the declarative model, specifically the covariances between the presented evidence and/or context nodes. However, the small sample of 20 scenarios represents a very small fraction of the declarative model's scenario space.[5] We minimized the potential for sampling from the declarative model to constrain the behavioral model by limiting the number of scenarios sampled in this manner and by including a second set of scenarios where node states were selected by the experimenter. Furthermore, pilots were kept naive with regard to the structure, conditional probabilities, and inferences of the declarative model. That the behavioral model differed substantially in structure from the declarative model indicates that sampling scenarios from the declarative model did not overly constrain pilot inferences in the simulation.

Pilot behavior was best predicted by a much simpler model in which ID was inferred from only three attributes (IFF, group posture, and group origin). This suggests that the set of variables that was identified during elicitation to be important in the inference of ID may be different to that which influences skilled behavior. Although pilots did revise their beliefs when presented with evidence from other sources, they did not do so in a consistent manner. It is not clear whether individual pilots were inconsistent in their use of the other information sources, or whether use was inconsistent across pilots. With regard to the former possibility, the differences between the declarative and behavioral models may reflect a difference in cognitive load between interview and simulation. In an interview, pilots may take time to consider all of the relevant variables and are able to use artifacts such as pen and paper to reduce working memory load (Hutchins, 1995). In simulation or flight, pilots must keep information pertaining to a task in working memory,

which imposes a limit to the number of variables that may be considered (Miller, 1956). That the behavioral model included four variables (including ID) is consistent with other research showing a limit of four variables in the information processing of complex tasks such as the interpretation of statistical interactions (Halford, Baker, McCredden, & Bain, 2005).

The design of the present task required pilots to make an inference of ID following the acquisition of each piece of evidence. This sequential sample/respond design allows all evidence accrued prior to a given search step to be summarized in a prior (to this step) belief of ID. On each step, it is only necessary to consider this prior belief of ID and the LR of the new evidence, thereby reducing working memory load. It is not clear whether pilots adopted this strategy, or whether they maintained representations of prior evidence in memory that continued to influence their inference.

The best model of pilot inference was one based on three attributes of ID: IFF, group origin, and group posture. However, this combination of attributes was not sufficiently diagnostic to fully discriminate between cognitive categories of friendly and hostile ID. In practice, military personnel perform ID of entities using legally binding decision trees (rules of engagement; ROE) that define necessary conditions for intervention. It is likely that pilot inference in the present study was guided by considerations of ROE because the attributes captured by the behavioral model and those considered in instances of non-compensatory search comprise a subset of conditions normally associated with hostile intent; that is, the adoption of an attack profile, the targeting of weapons, and the failure to respond to verbal and/or electronic queries, for example, from IFF (Cole, Drew, McLaughlin, & Mandsager, 2009). In addition to the modeling of pilot behavior, BNs may also prove to be useful in refining ROE, provided sufficient data are available to generate normative models of real-world processes.

A hypothesis of this study was that people generate causal mental models of generative processes in the world that support predictive and diagnostic inference (Sloman, 2005; Sloman & Lagnado, 2015). However, the problem of inferring ID from sources of evidence may alternatively be considered a problem of categorization on the basis of a number of attributes. Psychological models of categorization are based on similarity to exemplars of categories in memory (Anderson, 1991) and depend on associative rather than causal relationships (see Kruschke, 2008; for a review). Exemplar-based categorization may be modeled by naïve Bayes classifiers, where the category node is directly causal of all attribute nodes (Danks, 2014). The naïve Bayes architecture of the behavioral model is, therefore, consistent with the possibility that pilots performed inference of ID as a task of associative categorization, rather than one of causal inference. Therefore, although causal models were found to be successful in predicting pilot inference of ID, we cannot conclude that their mental models for this task were causal.

While we cannot conclude that inference about ID was based on causal reasoning, it remains likely that prediction and decision planning are based on predictive and counterfactual reasoning using causal models. In other words, it is possible that while inferences relating to questions of "who" or "what" may be made associatively, those relating to questions of "how," "why," "what next," and "what if" may require causal knowledge.

Further research is required to examine whether these other inferences, for example prediction of the consequences of planned actions as in naturalistic decision making (Klein, 1998), are supported by causal mental models.

The judgment and decision-making literature have a long history of evaluating the rationality of human inference. People commonly depart from logical or statistical accuracy in their judgments, instead demonstrating a range of biases and heuristics (e.g., Tversky & Kahneman, 1974). While these have been interpreted to reflect limitations of judgment, more recent research couches heuristics as potentially adaptive when environments are uncertain (Gigerenzer & Gaissmaier, 2011). There is also a growing body of literature that suggests that, on average, human judgment *does* conform to rational statistical principles (Gigerenzer & Hoffrage, 1995; Oaksford & Chater, 2007; Griffiths, Kemp, & Tenenbaum, 2008; but see Jones & Love, 2011; Rottman & Hastie, 2014). We do not assert that the pilots were Bayesian reasoners. Indeed, the common failures to revise prior beliefs on the basis of context and evidence represent neglect of the base rates (Tversky & Kahneman, 1974) and conservatism of belief updating (Phillips & Edwards, 1966). Nevertheless, pilot beliefs were well described by the Bayesian behavioral model. While this model is a Bayesian model, it makes no claims about the optimality of pilot behavior with respect to the world, given it was not populated with real-world data. It, thus, represents a descriptive model rather than a normative or rational one (see Tauber, Navarro, Perfors, & Steyvers, 2017).

Gigerenzer and Todd (1999) have reported that people commonly do not consider all available information when making decisions but instead adopt "fast and frugal heuristics" where search is based on only a subset of the available information. This is particularly so under conditions of time pressure (Rieskamp & Hoffrage, 2008) or when the cost of gathering information is high (Bröder, 2003). If the information is considered in order of decreasing validity, that is, the probability that the cue identifies the correct decision, the resulting decision may be close to optimal and heuristics developed on small data sets generalize well to new data (Gigerenzer & Todd, 1999). In the present study, pilots generally searched information sources in the order of decreasing utility, as reflected by mutual information with ID. This is generally consistent with the direction of search by cue validity (Gigerenzer & Todd, 1999) or cue success (Newell, Rakow, Weston, & Shanks, 2004). Search by pilots resulted in a near-optimal rate of reduction in the uncertainty of ID. However, in contrast to fast and frugal heuristics, search was generally found to be compensatory. That is, beliefs seldom reached certainty during the course of a scenario and pilots often continued to revise their beliefs until all available evidence had been exhausted. It should be noted that while pilots in this study were instructed to perform an exhaustive search, they were free to revise or to not revise their beliefs at each search step. Further research is required to determine whether pilots adopt a noncompensatory strategy when given the option to self-terminate search.

As cited in the Introduction, BNs may provide aids to human inference where good real-world data exist (e.g., Bladon et al., 2002; Das et al., 2002). Bayesian models of expert behavior may also prove to be useful in training for good inference. While declarative domain knowledge may be culturally transmitted by explicit instruction, implicit

knowledge must be gained by experience. It may be possible to represent the domain knowledge of experts in a BN and to train that BN on their behavioral inferences. The resulting model could then be used to train less-experienced people, particularly in cases where real-world experience is limited or costly to obtain.

The pilots in this study, including those from whom the declarative model was elicited, had a range of expertise from 200 to 2500 h on type. It is of interest to determine whether there are systematic differences in knowledge and inference between experts and less-experienced pilots. Unfortunately, the complexity of these scenarios required substantial data to develop robust models, and there were insufficient data at the level of the individual pilot in this study to be able to perform robust comparisons across individuals. This is the topic of ongoing research.

In conclusion, the present study demonstrates the potential of Bayesian networks to provide a unified and coherent framework for the study of expert knowledge, information search, and inference in complex environments. The framework also lends itself to the study of prediction and decision making in complex environments, although these were not examined in the present study.

## Acknowledgments

We thank the pilots from RAAF Squadrons 3 and 77 for participating in this study, and Russell Martin, Geoff Stuart, and three anonymous reviewers for helpful comments.

## Notes

1. For comparison, the KL divergence of a null model that generated random responses that were resampled from the distribution of pilot responses was 0.74.
2. We also examined models that allowed for (a) individual differences in the relationship between ID and each of the evidence nodes, and (b) interactions between the evidence nodes, by reversing the direction of each of the arcs from ID to the evidence nodes. This impaired the predictive ability of the model.
3. The pilot node was also sampled in this process.
4. The entropy of ID is the MI of ID with itself.
5. The small number of scenarios may have constrained the behavioral model, but extensive sampling of scenarios was not practical.

## References

Adams, S. (1998). Practical considerations for measuring situational awareness. In *Proceedings for the Third Annual Symposium on Situational Awareness in the Tactical Air Environment* (pp. 157–164). Piney Point, MD: Naval Air Warfare Center.

Anderson, J. R. (1991). The adaptive nature of human cognition. *Psychological Review*, *98*, 409–429.

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions*, *53*, 370–418.

Bladon, P., Hall, R. J., & Wright, W. A. (2002). Situation assessment using graphical models. In *Proceedings of the 5th International Conference on Information Fusion* (pp. 886–893). Annapolis, MD: IEEE.

Boyd, J. R. (1995). The essence of winning and losing. A 5-slide presentation. Available at http://www.danford.net/boyd/essence.htm. Accessed June 1, 2017.

Bröder, A. (2003). Decision making with the "adaptive toolbox": Influence of environmental structure, intelligence and working memory load. *Journal of Experimental Psychology: Learning*, *29*, 611–625.

Cole, A., Drew, P., McLaughlin, R., & Mandsager, D. (2009). *Rules of engagement handbook*. Sanremo, Italy: International Institute of Humanitarian Law.

Cousineau, D. (2005). Confidence intervals in within-subject designs: A simpler solution to Loftus and Masson's method. *Tutorials in Quantitative Methods for Psychology*, *1*, 42–45.

Danks, D. (2014). *Unifying the mind*. Cambridge, MA: The MIT Press.

Das, S., Grey, R., & Gonsalves, P. (2002). Situation assessment via Bayesian belief networks. In *Proceedings of the 5th International Conference on Information Fusion* (pp. 664–671). Annapolis, MD: IEEE.

Directorate of Defence Aviation and Air Force Safety (2010). Situational awareness. *Aviation Safety Spotlight*, *4*, 2–3.

Durso, F. T., Truitt, T. R., Hackworth, C. A., Crutchfield, J. M., Nikolic, D., Moertl, P. M., Ohrt, D., & Manning, C. A. (1995). Expertise and chess: A pilot study comparing situation awareness methodologies. In D. J. Garland & M. Endsley (Eds.), *Experimental Analysis and Measurement of Situation Awareness*. Embry-Riddle Aeronautical University Press.

Endsley, M. (1995a). Toward a theory of situation awareness in dynamic systems. *Human Factors*, *37*, 32–64.

Endsley, M. (1995b). Measurement of situation awareness in dynamic systems. *Human Factors*, *37*, 65–84.

Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic decision making. *Annual Review of Psychology*, *62*, 451–482.

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instructions: Frequency formats. *Psychological Review*, *102*, 684–704.

Gigerenzer, G., & Todd, P.M., & ABC Research Group (1999). *Simple heuristics that make us smart*. Oxford, UK: Oxford University Press.

Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., Kushnir, T., & Danks, D. (2004). A theory of causal learning in children: Causal maps and Bayes nets. *Psychological Review*, *111*, 3–32.

Griffiths, T. L., Kemp, C., & Tenenbaum, J. B. (2008). Bayesian models of cognition. In R. Sun (Ed.), *The Cambridge handbook of computational psychology*. New York: Cambridge University Press.

Hagmayer, Y., & Sloman, S. A. (2009). Decision makers conceive of their choices as interventions. *Journal of Experimental Psychology: General*, *138*, 22–38.

Halford, G. S., Baker, R., McCredden, J. E., & Bain, J. D. (2005). How many variables can humans process? *Psychological Science*, *16*, 70–76.

Hutchins, E. (1995). How a cockpit remembers its speeds. *Cognitive Science*, *19*, 265–288.

Jones, M., & Love, B. C. (2011). Bayesian fundamentalism or enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. *Behaviour and Brain Sciences*, *34*, 169–231.

Klein, G. (1998). *Sources of power. How people make decisions*. Cambridge, MA: The MIT Press.

Korb, K. B., Hope, L. R., Nicholson, A. E., & Axnick, K. (2004). Varieties of causal intervention. In C. Zhang, H. W. Guesgen, & W. K. Yeap (Eds.), *PRICAI 2004: Trends in Artificial Intelligence. Proceedings of the 8th Pacific Rim International Conference on Artificial Intelligence* (pp. 322–331). Berlin: Springer.

Kruschke, J. R. (2008). Models of categorization. In R. Sun (Ed.), *The cambridge handbook of computational psychology* (pp. 267–301). New York: Cambridge University Press.

Marr, D. (1982). *Vision*. New York: W. H. Freeman and Company.
Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *63*, 81–97.
Mitchell, T. M. (1997). *Machine learning*. New York: McGraw-Hill.
Neapolitan, R. (1990). *Probabilistic reasoning in expert systems: Theory and algorithms*. New York: John Wiley & Sons.
Neisser, U. (1976). *Cognition and reality. Principles and implications of cognitive psychology*. New York: W. H. Freeman and Company.
Newell, B. R., Rakow, T., Weston, N. J., & Shanks, D. R. (2004). Search strategies in decision making: The success of "success". *Journal of Behavioural Decision Making*, *17*, 117–137.
Oaksford, M., & Chater, N. (2007). *Bayesian rationality: The probabilistic approach to human reasoning*. Oxford: Oxford University Press.
Pearl, J. (2009). *Causality. Models, reasoning, and inference* (2nd ed.). New York: Cambridge University Press.
Phillips, L. D., & Edwards, W. (1966). Conservatism in a simple probability inference task. *Journal of Experimental Psychology*, *72*, 346–354.
Rieskamp, J., & Hoffrage, U. (2008). Inferences under time pressure: How opportunity costs affect strategy selection. *Acta Psychologica*, *127*, 258–276.
Rottman, B. M., & Hastie, R. (2014). Reasoning about causal relationships: Inferences on causal networks. *Psychological Bulletin*, *140*, 109–139.
Shlens, J. (2014). Notes on Kullback-Leibler divergence and likelihood theory. Preprint. Available at arXiv:1404:2000v1 [cs.IT]. Accessed February 17, 2016.
Sloman, S. (2005). *Causal models. How people think about the world and its alternatives*. New York: Oxford University Press.
Sloman, S. A., & Hagmayer, Y. (2006). The causal psycho-logic of choice. *Trends in Cognitive Sciences*, *10*, 407–412.
Sloman, S. A., & Lagnado, D. (2015). Causality in thought. *Annual Review of Psychology*, *66*, 223–247.
Tauber, S., Navarro, D. J., Perfors, A., & Steyvers, M. (2017). Bayesian models of cognition revisited: Setting optimality aside and letting data drive psychological theory. *Psychological Review*, *124*, 410–441.
Taylor, R. M. (1990). Situational awareness rating technique (SART): The development of a tool for aircrew systems design. In *Proceedings of the AGARD AMP symposium on Situational awareness in aerospace operations* (pp. 3/1-3/17), (AGARD-CP-478). Neuilly-sur-Seine: NATO AGARD.
Tversky, A., & Kahneman, D. (1974). Judgement under uncertainty: Heuristics and biases. *Science*, *185*, 1124–1131.