Methods for Representing Bias in Bayesian Networks

Eric Carlson, Sean Guarino, Jonathan Pfautz
Charles River Analytics, Inc.
625 Mount Auburn Street
Cambridge, MA 02138

Abstract

Bias is intrinsic to observation and reasoning in both humans and automated systems. Bayesian Belief Networks (BBNs) are well suited for representing these biases and for applying bias models to improve reasoning practices, but there are a number of different ways that bias can be represented and integrated into reasoning processes using BBNs. In this paper, we describe a number of methods to model biases using BBNs and discuss the strengths and weaknesses of each method.

1. INTRODUCTION

Bias is intrinsic to observation and reasoning. Though the concept carries connotations of human judgment, bias also applies to automated systems, introduced by the limitations of their capabilities. Reasoning about information that includes bias (i.e., processed information, whether from human of machine) requires reasoning about the information, itself, and about the biases that influenced it. Humans do this naturally. In rich human-to-human interactions, each person derives an understanding of the biases involved from shared context and estimates of the other's attitudes and beliefs. In other circumstances, such as shallow person-to-person interactions (e.g., reading a restaurant review from an unknown person) or interactions involving automated processes (e.g., getting directions from a GPS; incorporating human reports into an automated decision aide; integrating contributions from multiple sensor systems in a data fusion system), biases and their influences need to be made explicit. As Hastie & Dawes (2001) argue, incorporating an explicit model of biases and their influences into reasoning processes can lead to more robust and accurate reasoning in both humans and automated systems.

Bayesian Belief Networks (BBNs) are well suited for modeling biases in automated processing systems and decision aides. Many factors contribute to bias, interacting in a complex manner with each other and with the overall bias. BBNs represent the type of probabilistic influences and causal relationships required to capture this interaction (Pearl & Russell, 2000; Pearl, 2001). Furthermore, the graphical nature of BBNs further supports the expression of these relationships by providing an intuitive method to capture contributing factors and influences. In addition to providing an applicable modeling approach for capturing biases, BBNs are already applied in many fields where consideration of biases has the potential to make significant contributions to performance and realism, such as military intelligence (Koelle et al., 2006; Pfautz et al., 2005a; Pfautz et al., 2005b), medical diagnostics (Kononenko, 1993; Parmigiani, 2002; Nikovski, 2000), and human behavior modeling (Guarino et al., 2006; Hudlicka & Pfautz, 2002; Neal Reilly et al., 2007; Pfautz & Lovell, 2008).

To advance the incorporation of bias models in these fields and others, in this paper we discuss the role of biases in the decision making process (which includes, for our purposes here, observation, reasoning, and decision selection), several ways bias can be modeled using BBNs, and the benefits and drawbacks of each of these methods.

2. BACKGROUND

The study of biases to date largely focuses on cognitive biases. Several attempts have been made to categorize different types of bias and to identify how they affect the decision-making process. One method for classification is to look at the source of the bias, for instance, dividing uncertainty into forms that come from computational models as opposed to human interpretation (Schunn, Kirschenbaum, & Trafton, 2003). Another method is to examine the use of bias and uncertainty in the decision-making process, resulting in categories, which has resulted in categories such as *executional uncertainty*, *goal uncertainty*, and *environmental uncertainty* (Yovits & Abilock, 1974). Another set of classifications developed by Lipshitz and Strauss (1996) divides forms of uncertainty into *inadequate understanding*, *lack of information*, and *conflicted alternatives.* Similar taxonomies were developed by Schunn et al. (2003) and Klein (1998). These taxonomies can prove to be useful in attempts to develop descriptive models of human reasoning. For example, Lipshitz & Straus (1996) discuss

five strategies for reasoning under uncertainty: 1) reduce uncertainty by collecting more information; 2) use assumptions to fill in gaps of knowledge; 3) weigh pros and cons; 4) forestall; and 5) suppress uncertain information. While these classifications of uncertainty and an understanding of biases they introduce to decision-making have been useful in the development of models of human behavior, they may not generalize to other types of biases.

3. ROLE OF BIAS

For the purpose of incorporating consideration of bias into reasoning process, we are concerned with bias in two separate roles. First, because bias impacts the creation of the products of observations and reasoning processes, it must be accounted for in the *interpretation* of those products. Limitations, methods, and, in the case of humans, preferences and cognitive biases introduce a systematic modification into an observed product. This modification must be identified and defined to properly reason based on these products. Elaborating on the earlier example, consider a negative review of a French restaurant written by someone who dislikes French food. Whether he is cognizant of this influence or not, the product of his observation—the review—incorporates his pre-existing preference. To reason based on this review, anyone reading it needs to recognize and correct for the preferences of the reviewer. Automated systems may not have personal preferences, but their technical limitations can introduce similar biases. Consider a sensor that detects the presence of humans based on heat signatures. Because readings are based on the contrast between the person and the ambient temperature, this sensor has a higher occurrence of false negatives when the temperature is above body temperature. So, a reading showing no people present on a 100°F day may be disingenuous because it is the product of both the reading and the hidden bias introduced by its technical limitations. As with the previous example of human bias, the consumer of this automated report—human or automated system—must reason about both the contributing bias and the information, itself, to accurately use the product.

Second, bias impacts the *reasoning* process applied to make decisions based on information products. The consumer introduces its own systematic modification of the information based both on its own biases and on the perceived biases incorporated in the product. For example, the analysis system using reports from the heat sensor may incorporate the fact that it does not function if the temperature is over 100°F , and disregard the sensor's information products on a particularly hot days. Similarly, the analysis may favor one sensor type over another for gathering specific information, regardless of specific conditions (e.g., an analysis system may trust a radar over an eye witness due to a bias against non-technical sources). In this role, bias is not considered solely in the context of information production (though this may be considered); these biases consider how the information is being used and the reasoning processes involved.

These two roles are cyclic, as the results of a reasoning process can be viewed as its own information product. If there are known biases in that product, an estimate of those biases may become an element in a new consumer's reasoning processes, alongside other reasoning biases of the consumer. When the information product being interpreted pertains to an observable truth (e.g., a sensor detecting some object), understanding the influence of bias allows the consumer to determine the accuracy of the product and to integrate that accuracy information into its own reasoning processes. When the product pertains to a subjective belief or assessment (e.g., an opinion about a restaurant), understanding the contributing biases allows the consumer to determine how to integrate those biases with its own biases.

These two roles comprise use cases for bias models, each with their own concerns motivating different design decisions. In the *interpretation* role, a model of bias can serve as a mechanism to correct for biases. Here, the details of the sources of those biases may not necessarily be important. Rather, it is important to correct for errors caused by biases. In the *reasoning* role, a model can be used to self-regulate against the introduction of additional biases, as well as to increase the accuracy of the consumer's estimation of biases contributing to a product, which allows information to be incorporated into the consumers own reasoning at the highest fidelity possible. Here, the details of the sources of those biases may be extremely important, as different meta-information and information may have a direct influence in the reasoning process.

4. THE STRUCTURE OF BIAS

As a concept, bias is closely related to meta-information. Meta-information is information about information. That is, information that serves to qualify and give context to other information. For example, if a sensor reading is information, the fact "the reading is two weeks old" is meta-information—information about the report. For a more extensive discussion of meta-information, see (Guarino et al., 2006). Whereas meta-information is a statement of fact ("the report is old"), bias is the effect meta-information has on observations and reasoning processes ("because the report is old, its contents are probably inaccurate"). Thus, information types can be divided into three levels:

1) the information, itself (e.g., the contents of the report)
2) meta-information (e.g., information about the report)
3) biases (e.g., the impact information *about* the report—the meta-information—has on the information *in* the report)

Biases are derived from meta-information by combining that meta-information with elements of the information.

For example, a two week old sensor reading showing the location of people in an open setting would not convey their current location with high confidence, while a two week old sensor reading showing the location of buildings would represent their current position with a high degree of certainty. So, in this example, the bias ("the information in the reading is wildly inaccurate") is derived from a factor of the information ("people move frequently") combined with meta-information ("the report is ten days old"). This same logic holds for subjective assessments. In the restaurant review example,

- Meta-information: The reviewer hates French food
- Information: The restaurant is French
- Bias: The reviewer was predisposed to hate the restaurant, regardless of its quality

These definitions of information types and the derivation of bias are the basis for the structure of our bias models.

5. BIAS MODELS

In this section, we present a number of ways to model bias, and we discuss the advantages and disadvantages of each model in light of the roles of bias (see section 3) and additional concerns about model use and creation. Bias models vary along two dimensions: the level of detail expressed about the bias and the level of integration with the reasoning model to which it is meant to contribute.

5.1 IMPLICIT BIAS MODEL

The implicit bias model does not contain a representation of the bias in its structure. Instead bias is expressed in the relationship between nodes of the existing elements of the model. Insomuch as it exists anywhere, the bias exists in each node's Conditional Probability Tables (CPTs). The effect this bias exerts on the product of the model—the observation, decision, behavior, etc.—is a change in the beliefs of the nodes. The bias, itself, is not explicitly represented separate from the state information of the model. For example, see Figure 5-1, an implicit bias model of our previous heat sensor example.

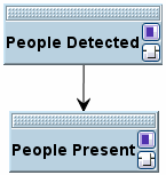


Figure 5-1: Implicit bias model structure of the heat sensor example. Bias is represented only in the CPTs.

The sole factor represented as contributing to whether people are present is the number of people detected by the sensor. The bias in this model is expressed as uncertainty in the outcome. For positive readings, the likelihood of people being present is high. Because there are conditions that can increase the likelihood of false negatives, though, a negative reading leads to a lower certainty of people not being present (see Figure 5-2).

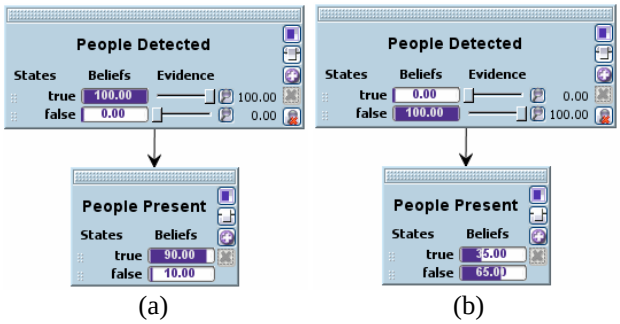


Figure 5-2: (a) left, shows the high belief that people are present based on a positive reading of the heat-based sensor; (b), right, shows a less certain belief that people are not present based on a negative reading of the same sensor. The bias is reflected in the increased uncertainty due to the possibilities of false negatives.

The implicit bias model reflects the simplest case. Though it does reflect the reality of the situation, this model is insufficient in most other ways. Because elements that contribute to the bias (i.e., meta-information) are not explicitly represented, the bias is reflected in a permanent change in confidence rather than reflecting specific conditions (e.g., because the ambient temperature is not explicitly represented, the confidence cannot change based on the specific value of that variable). Instead, this model merely represents that bias is possible in the reasoning process. This model may be sufficient for representing bias while interpreting data because the value of the relevant meta-information may not be available to the consumer. However, because it does not explicitly describe the contributing factors and applies the bias as a consistent change in certainty rather than on a case-by-case basis, it is ineffective at providing a nuanced bias model for reasoning.

5.2 INTEGRATED BIAS MODEL

In an integrated model, the factors that contribute to bias (i.e., meta-information) are explicitly represented as nodes in the network and are fully integrated into the model of the observation, reasoning process, behavior, etc. The bias—the effect of this meta-information—is still contained in the CPTs. Like the implicit model, there is a bias in the computational process, but that bias is not explicitly represented as a node in the BBN. Figure 5-3 expands Figure 5-1 into an integrated model by adding Ambient Temperature as an input node.

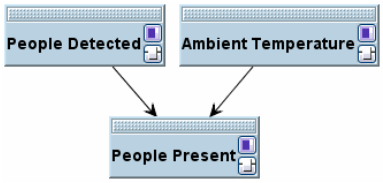
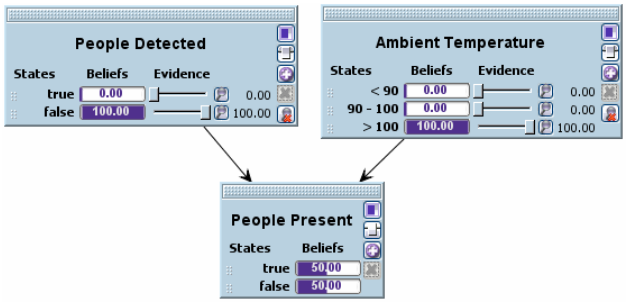
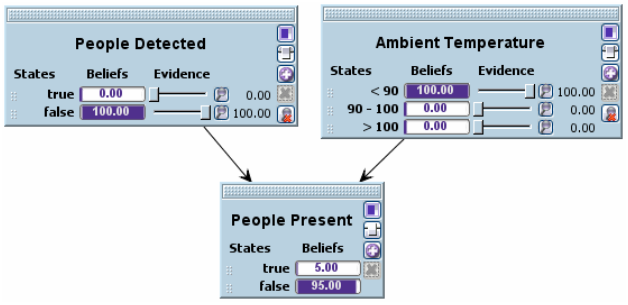


Figure 5-3: An integrated bias model of the heat sensor example. Meta-informational factors are represented. Bias is represented in the CPTs.

This inclusion of factors that moderate biases allows the bias model to account for the exact value of relevant meta-information, allowing the bias to change dynamically (see Figure 5-4). Furthermore, because each factor is expressed independently, their combined effect on the reasoning process can be nuanced.



(a)



(b)

Figure 5-4: Integrated bias model of the heat sensor example. In (a), a high ambient temperature increases uncertainty. In (b), a low ambient temperature decreases uncertainty. The bias reacts in real-time to conditions, increasing accuracy of the model.

In an integrated bias model, factors contributing to bias are explicitly expressed, so these models are more accurate, and, therefore, better than implicit models in an interpretation role. However, as in the implicit model, the effect of these factors is still captured fully in the CPTs. For this reason, expansion of the model is difficult, as additions could require significant modifications to those CPTs. Therefore, in a reasoning role it is difficult to adapt parts of an integrated bias model for reuse in a larger reasoning model.

5.3 CONSOLIDATED UNKNOWN BIAS MODEL

In a consolidated unknown bias model, bias is expressed as a single node in the network, with connections to each of the nodes in the network. This single node is a "black box" meant to represent the amount of bias in the model with no concern for the cause of the bias (note: this node could be a placeholder for bias calculated using the standalone bias model discussed in sections 5.5 and 5.6). For an example of a consolidated unknown bias model, see Figure 5-5.

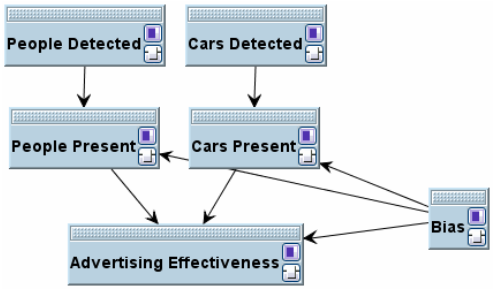


Figure 5-5: A consolidated unknown bias model, where the strength of the bias present is represented by a single node, which connects to all elements of the reasoning model.

This model does contain a mechanism to express bias in every part of the model, but it makes a large assumption about the distribution of that bias. The effect bias has on each element is expressed in the CPTs, which means that the specific effects of the bias strength is individual to each node, but the strength is shared. This model does represent the effect of bias on a gross level, so it can be used somewhat in an interpretation role, albeit with lower fidelity since all biases are expressed in a single dimension. The effect of the bias is hidden in the CPTs, and the factors that contribute to the bias are completely unstated, so in a reasoning role biases cannot be utilized by addition elements of a reasoning model.

5.4 DISTRIBUTED UNKNOWN BIAS MODEL

The distributed unknown bias model represents bias as a number of "black boxes", each having an effect on one or more elements of the reasoning model. Again, as black boxes, the factors contributing to each bias are not explicit. Bias nodes provide an overall representation of the biases in the reasoning components to which they are attached. For an example, see Figure 5-6.

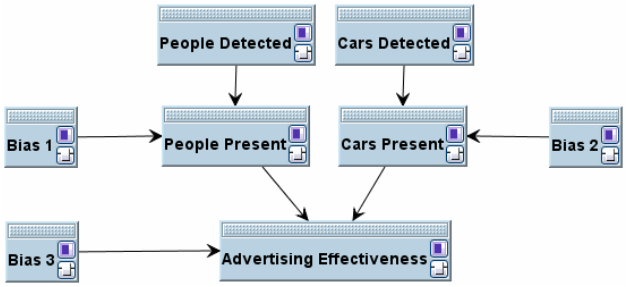


Figure 5-6: A distributed unknown bias model, where bias is represented as a number of unknowns, each connected to elements of the reasoning model.

Distributed unknown bias models are superior to consolidated unknown bias models because they express a more nuanced situation reflecting the susceptibility of various elements of the model to different biases. The bias nodes play a similar role to meta-information nodes in an integrated bias model, but, as black-box bias modules, they consolidate all factors contributing to a particular bias into a single node. In an interpretation role, these models are more useful than implicit bias models because at least some gauge of the strength of bias active in each element is present. However, unlike the integrated bias model, the meta-information factors that affect their strength are unknown. This reduces the already limited ability of bias factors in distributed unknown bias models to be integrated into an external reasoning model. Unlike the models representing meta-information factors explicitly, the ability to add factors is not a concern because they are aggregated together in a single node, so no CPTs need to be changed. However, without expressing the composition of the bias, the bias strengths and relationships are highly subjective.

5.5 STANDALONE BIAS MODEL

A standalone bias model expresses bias in an independent model separate from the reasoning model. This is distinct from the integrated model where factors are represented but are integrated with the reasoning model itself. The measure of bias resulting from this model can then be applied to the reasoning model, filling the black-box need of the consolidated or distributed bias model, or used alone. Bias is expressed explicitly as a single node. Each element in a reasoning model where bias is a factor would require an independent bias model. The mechanism by which each factor contributes to bias is hidden in the CPTs. For an example of a standalone bias model, in the heat-based human detector the meta-information factor “Ambient Temperature” could be expressed (alongside any other relevant factors) as explicit nodes. The effect that each factor has (i.e., that high temperature increases the uncertainty of negative readings) is still expressed only in the CPTs. This example is depicted in Figure 5-7.

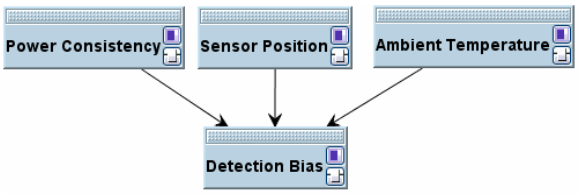


Figure 5-7: A standalone bias model of detection bias for a heat-based person detector. The product of this standalone model could then be applied in a reasoning model.

Like the integrated model, because standalone bias models represent the contribution of each of a set of factors to a bias explicitly, these models can dynamically capture bias, providing greater accuracy. Expressing factors in a separate model allows them to easily be applied as a factor in a large or frequently changing model. For this reason, standalone bias models excel in circumstances where a bias model might be applied independently at multiple points in a reasoning process.

For example, consider a data fusion application that receives sporadic inputs from a host of sensors. Rather than use a single monolithic model that integrates information from all sensors, standalone bias models could be used to dynamically assemble a model that represents only those sensors that are currently active. Because the majority of the sensors are silent at any given time, this improves the efficiency of bias application in such conditions. However, this autonomy has a tradeoff in that bias is consolidated into a single metric resulting in the influence of specific pieces of meta-information having limited nuance in their effect on the reasoning process. Furthermore, an element or even a network fragment might be repeated in multiple standalone models leading to wasteful repetitive computation. Nevertheless, due to the explicit representation of meta-informational factors and simple portability, this type of model applies well in both interpretation and reasoning roles.

5.6 FULLY ENUMERATED STANDALONE BIAS MODEL

Fully enumerated standalone bias models explicitly represent both the meta-information that causes the bias and the element that defines how that meta-information contributes to bias (as discussed in Section 4). Rather than a single model for each bias type as with the standalone bias model, fully enumerated standalone bias model have a single model for each element of the information that, when paired with meta-information, could introduce a bias. These models express all factors contributing to bias and the bias itself as elements in the network, rather than being contained in the CPTs. For an example of fully enumerated standalone bias models, see Figure 5-8.

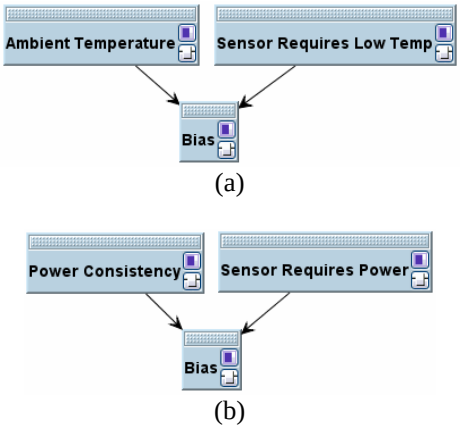


Figure 5-8: Fully enumerated standalone bias models for (a) bias related sensors whose performance is affected by temperature, and (b) bias related to sensors whose performance is affected by power supply.

Similar to the way standalone bias model can be applied dynamically based on the biases present, fully enumerated standalone bias models can be applied based on the definition of the system creating the product. So, a system using these needs a model for each possible property of the data sources. It can then apply them based on the definition of each source. For example, in a fusion system designed to dynamically calculate bias for any configuration of sensors, a bias model could be automatically assembled for each sensor based on the operating characteristics of that sensor. The heat sensor, defined as requiring low temperature, would incorporate biases related to that requirement. Because these networks determine the bias introduced by each factor separately, their integration into a reasoning process can be more nuanced than representations that consolidate bias into a single measure. This, along with the transparency of contributing factors, makes them ideal in a reasoning role.

6. CONCLUSIONS

There are numerous ways to represent bias as a BBN, each of which has its own strengths and weaknesses. Models of bias provide a mechanism to correct for bias to increase accuracy and to integrate biased information into human and automated reasoning processes. The most advantageous form of model for a particular situation depends on its intended use.

By systematically examining the composition of bias, we have identified factors in its composition. The various model types we discussed make use of this definition by incorporating various factors at a range of fidelities, making specific elements more or less accessible. Additionally, we have defined two separate roles bias can play in reasoning processes. These roles form the basis for use cases, which we have used to evaluate each of the types of models. Of the models discussed, the more nuanced the application of bias to elements that contributed to the production of information, the greater the benefit in accurately interpreting the product of reasoning processes without introducing additional biases. To reason based on those products, those models that include the greatest level of detail and autonomy for factors that contribute to bias can be more easily and accurately integrated into reasoning processes.

7. DISCUSSION

This set of bias representations encapsulates a significant range of capabilities and tradeoffs. Among the most prominent difference between these representations is the degree of specificity about the sources of bias. In certain applications, like accounting for bias from a technical sensor, these bias factors can be easily identified and described. In others, like accounting for bias in human reasoning, these sources are obscured and can only be hypothesized through intense effort, and are largely unverifiable. In light of these impediments, going forward we need to determine what guidelines could be established to govern the applicability of different styles. How can uncertainty about the causes of bias be mitigated? Is there a way to create representations that don't incorporate unspecified sources of bias, but that are applicable in situations where those sources are vaguely defined? Or, are there ways to use black box bias measures without fully sacrificing the attribution that identifying specific sources provides? Is this attribution of bias to particular sources necessarily important (e.g., for accountability, trust)? What conditions of use make attribution important (e.g., frequent updates, logic exposed to the user)? The complexity of specificity results, too, in a gain in precision in the end bias measure. Can factors contributing to bias be calculated precisely enough to warrant this precision in the end product?

Acknowledgements

The authors would like to express their deepest gratitude to the subject matter experts who contributed to our understanding of biases. Additionally, we would like to thank Dr. Greg Zacharias for funding our work on Bayesian Belief Networks.

References

Guarino, S., Pfautz, J., Cox, Z., & Roth, E. (2006). Modeling Human Reasoning About Meta-Information. In *Proceedings of 4th Bayesian Modeling Applications Workshop at the 22nd Annual Conference on Uncertainty in AI: UAI '06*. Cambridge, Massachusetts.

Hastie, R. & Dawes, R. M. (2001). *Rational Choice in an Uncertain World: The Psychology of Judgment and Decision-Making*. London, UK: Sage Publications.

Hudlicka, E. & Pfautz, J. (2002). Architecture and Representation Requirements for Modeling Effects of Behavior Moderators. In *Proceedings of Proceedings of 11th Conference on Computer-Generated Forces - Behavior Representation*. Orlando, FL.

Klein, G. A. (1998). *Sources of Power: How People Make Decisions*. Cambridge, MA: MIT Press.

Koelle, D., Pfautz, J., Farry, M., Cox, Z., Catto, G., & Campolongo, J. (2006). Applications of Bayesian Belief Networks in Social Network Analysis. In *Proceedings of 4th Bayesian Modeling Applications Workshop at the 22nd Annual Conference on Uncertainty in AI: UAI '06*. Cambridge, Massachusetts.

Kononenko, I. (1993). Inductive And Bayesian Learning in Medical Diagnosis. *Applied Artificial Intelligence*, 7(4), 317-337.

Lipshitz, R. & Strauss, O. (1996). How Decision-Makers Cope With Uncertainty. In *Proceedings of Proceedings of the Human Factors and Ergonomics Society 40th Annual Meeting-1996*, (pp. 189-193). Philadelphia: Human Factors Society.

Neal Reilly, W. S., Bayley, C., Koelle, D., Marotta, S., Pfautz, J., & Keeney, M. (2007). *Culturally Aware Agents for Training Environments (CAATE): Final Report.* (Rep. No. R070101). Cambridge, MA: Charles River Analytics, Inc.

Nikovski, D. (2000). Constructing Bayesian Networks for Medical Diagnosis From Incomplete and Partially Correct Statistics. *IEEE Transactions on Knowledge and Data Engineering*, 12(4), 509-516.

Parmigiani, G. (2002). *Modeling in Medical Decision Making: A Bayesian Approach*. John Wiley and Sons.

Pearl, J. (2001). *Causality: Models, Reasoning, and Inference*. Cambridge Univ Press.

Pearl, J. & Russell, S. (2000). *Bayesian Networks.*

Pfautz, J., Fouse, A., Roth, E., & Karabaich, B. (2005a). Supporting Reasoning About Cultural and Organizational Influences in an Intelligence Analysis Decision Aid. In *Proceedings of International Conference on Intelligence Analysis*. McLean, VA.

Pfautz, J. & Lovell, S. (2008). Methods for the Analysis of Social and Organizational Aspects of the Work Domain. In A. Bisantz & C. Burns (Eds.), *Applications of Cognitive Work Analysis*. Lawrence Erlbaum Associates.

Pfautz, J., Roth, E., Bisantz, A., Fouse, A., Madden, S., & Fichtl, T. (2005b). The Impact of Meta-Information on Decision-Making in Intelligence Operations. In *Proceedings of Human Factors and Ergonomics Society Annual Meeting*. Orlando, FL.

Schunn, C. D., Kirschenbaum, S. S., & Trafton, J. G. (2003). The Ecology of Uncertainty: Sources, Indicators, and Strategies for Information Uncertainty. http://www.au.af.mil/au/awc/awcgate/navy/nrl_uncertainty_taxonomy.pdf [On-line].

Yovits, M. C. & Abilock, J. (1974). A Semiotic Framework for Information Science Leading to the Development of a Quantitative Measure of Information. In *Proceedings of 37th American Society for Information Sciences Meeting*, (pp. 163-168).