

Running head: CAUSALITY IN JUDGMENT

The Role of Causality in Judgment Under Uncertainty

Tevye R. Krynski & Joshua B. Tenenbaum

Department of Brain & Cognitive Sciences, Massachusetts Institute of Technology

77 Massachusetts Ave, Cambridge, MA 02139 USA

JBT was supported by the Paul E. Newton Chair and the James S. McDonnell Foundation Causal Learning Research Collaborative. TRK was supported by an NSF graduate student fellowship. We thank Brigid Dwyer and Laila Shabir for assistance in running experiments, Tom Griffiths for helpful discussions and suggestions about data analysis, and Elizabeth Bonawitz for helpful comments on a previous draft. Address correspondence to: Joshua B. Tenenbaum, Department of Brain & Cognitive Sciences, 46-4015, 77 Massachusetts Ave., Massachusetts Institute of Technology, Cambridge, MA 02139. Phone: (617) 452-2010. Email: jbt@mit.edu.

## Abstract

Leading accounts of judgment under uncertainty evaluate performance within purely statistical frameworks, holding people to the standards of classical Bayesian (Tversky & Kahneman, 1974) or frequentist (Gigerenzer & Hoffrage, 1995) norms. We argue that these frameworks have limited ability to explain the success and flexibility of people's real-world judgments, and propose an alternative normative framework based on Bayesian inferences over causal models. Deviations from traditional norms of judgment, such as "base-rate neglect", may then be explained in terms of a mismatch between the statistics given to people and the causal models they intuitively construct to support probabilistic reasoning. Four experiments show that when a clear mapping can be established from given statistics to the parameters of an intuitive causal model, people are more likely to use the statistics appropriately, and that when the classical and causal Bayesian norms differ in their prescriptions, people's judgments are more consistent with causal Bayesian norms.

### The Role of Causality in Judgment Under Uncertainty

Everywhere in life, people are faced with situations that require intuitive judgments of probability. How likely is it that this person is trustworthy? That this meeting will end on time? That this pain in my side is a sign of a serious disease? Survival and success in the world depend on making judgments that are as accurate as possible given the limited amount of information that is often available. To explain how people make judgments under uncertainty, researchers typically invoke a computational framework to clarify the kinds of inputs, computations, and outputs that they expect people to use during judgment. We can view human judgments as approximations (sometimes better, sometimes worse) to modes of reasoning within a rational computational framework, where a computation is “rational” to the extent that it provides adaptive value in real-world tasks and environments. However, there is more than one rational framework for judgment under uncertainty, and behavior that looks irrational under one framework may look rational under a different framework. Because of this, evidence of “error-prone” behavior as judged by one framework may alternatively be viewed as evidence that a different rational framework is appropriate.

This paper considers the question of which computational framework best explains people’s judgments under uncertainty. To answer this, we must consider what kinds of real-world tasks and environments people encounter, which frameworks are best suited to these environments (i.e., which we should take to be normative), and how well these frameworks predict people’s actual judgments under uncertainty (i.e., which framework offers the best descriptive model). We will propose that a causal Bayesian framework, in which Bayesian inferences are made over causal models, represents a more appropriate normative standard and a more accurate descriptive model than previous frameworks for judgment under uncertainty.

The plan of the paper is as follows. We first review previous accounts of judgment under uncertainty, followed by the arguments for why a causal Bayesian framework provides a better normative standard for human judgment. We then present four experiments supporting the causal Bayesian framework as a descriptive model of people's judgments. Our experiments focus on the framework's ability to explain when and why people exhibit base-rate neglect, a well-known judgment phenomenon that has often been taken as a violation of classical Bayesian norms. Specifically, we test the hypotheses that people's judgments can be explained as approximations to Bayesian inference over appropriate causal models, and that base-rate neglect often occurs when experimenter-provided statistics do not map clearly onto parameters of the causal model participants are likely to invoke. We conclude by discussing implications of the causal Bayesian framework for other phenomena in probabilistic reasoning, and for improving the teaching of statistical reasoning.

#### Statistical frameworks for judgment under uncertainty

Most previous accounts – whether arguing for or against human adherence to rationality – take some framework of statistical inference to be the normative standard (Anderson, 1990; Gigerenzer & Hoffrage, 1995; McKenzie, 2003; Oaksford & Chater, 1994; Peterson & Beach, 1967; Shepard, 1987; Tversky & Kahneman, 1974). Statistical inference frameworks generally approach the judgment of an uncertain variable, such as whether someone has a disease, by considering both the current data, such as the person's symptoms, as well as past co-occurrences of the data and the uncertain variable, such as previous cases of patients with the same symptoms and various diseases. Because these frameworks focus on observations rather than knowledge, beliefs about the causal relationships between variables does not play a role in inference.

Using statistical inference frameworks as a rational standard, several hypotheses have been advanced to describe how people make judgments under uncertainty. Early studies of judgment suggested that people behaved as “intuitive statisticians” (Peterson & Beach, 1967), because their judgments corresponded closely to classical Bayesian statistical norms, which were presumed rational. Classical Bayesian norms explain how prior beliefs may be updated rationally in light of new data, via Bayes’ rule. To judge  $P(H | D)$ , the probability of an uncertain hypothesis  $H$  given some data  $D$ , Bayes’ rule prescribes a rational answer, as long as one knows (1)  $P(H)$ , the prior degree of belief in  $H$ , and (2)  $P(D | H)$  and  $P(D | \neg H)$ , the data expected if  $H$  were true and if  $H$  were false:

$$P(H | D) = \frac{P(H)P(D | H)}{P(D)} \quad (1)$$

where  $P(D) = P(H)P(D | H) + P(\neg H)P(D | \neg H)$ .

The intuitive statistician hypothesis did not reign for long. It was not able to account for a rapidly accumulating body of experimental evidence that people reliably violate Bayesian norms (Ajzen, 1977; Bar-Hillel, 1980; Eddy, 1982; Lyon & Slovic, 1976; Nisbett & Borgida, 1975; Tversky & Kahneman, 1974). For example, consider the “mammogram problem”, a Bayesian diagnosis problem which even doctors commonly fail (Eddy, 1982). One well-tested version comes from Gigerenzer and Hoffrage (1995), adapted from Eddy (1982), in which participants were told that the probability of breast cancer for any woman getting a screening is 1%, that 80% of women with cancer receive a positive mammogram, and that 9.6% of women without cancer also receive a positive mammogram. Participants were then asked to judge the likelihood that a woman who receives a positive mammogram actually has cancer. Participants often give answers of 70%-90% (Eddy, 1982; Gigerenzer & Hoffrage, 1995), while Bayes’ theorem prescribes a

much lower probability of 7.8%. In this case,  $H$  is “patient X has breast cancer”,  $D$  is “patient X received a positive mammogram”, and the required task is to judge  $P(H | D)$ , the probability that the patient has breast cancer given that she received a positive mammogram:

$$P(H | D) = \frac{P(H)P(D | H)}{P(H)P(D | H) + P(\neg H)P(D | \neg H)} = \frac{1\% \times 80\%}{1\% \times 80\% + 99\% \times 9.6\%} = 7.8\% \quad (2)$$

Kahneman and Tversky (1973) characterized the source of such errors as “neglect” of the base rate (in this case, the rate of cancer), which should be used to set  $P(H)$  in the above calculation of  $P(H | D)$  (in this case, the probability of cancer given a positive mammogram).<sup>1</sup>

The heuristics and biases view, which came to replace the intuitive statistician framework, sought to understand probabilistic judgments as heuristics, which approximate normative Bayesian statistical methods in many cases, but lead to systematic errors in others (Tversky & Kahneman, 1974). Given the focus of the heuristics and biases program on judgment errors, many concluded that people were ill-equipped to reason successfully under uncertainty. Slovic, Fischhoff, and Lichtenstein (1976) wrote: “It appears that people lack the correct programs for many important judgmental tasks.... it may be argued that we have not had the opportunity to evolve an intellect capable of dealing conceptually with uncertainty” (p. 174). Yet by the standards of engineered artificial intelligence systems, the human capacity for judgment under uncertainty is prodigious. People, but not computers, routinely make successful uncertain inferences on a wide and flexible range of complex real-world tasks. As Glymour (2001) memorably asked, “If we’re so dumb, how come we’re so smart?” (p. 8). Research in the heuristics and biases tradition generally did not address this question in a satisfying way.

These previous paradigms for analyzing judgment under uncertainty, including the Heuristics & Biases program (Tversky & Kahneman, 1974) and the Natural Frequency view (Gigerenzer & Hoffrage, 1995), have one commitment in common: they accept the

appropriateness of traditional statistical inference as a rational standard for human judgment. Purely statistical methods are best suited to reasoning about a small number of variables based on many observations of their patterns of co-occurrence – the typical situation in ideally controlled scientific experiments. In contrast, real-world reasoning typically involves the opposite scenario: complex systems with many relevant variables and a relatively small number of opportunities for observing their co-occurrences. Because of this complexity, the amount of data required for reliable inference with purely statistical frameworks, which generally grows exponentially in the number of variables, is often not available in real-world environments. The conventional statistical paradigms developed for idealized scientific inquiry may thus be inappropriate as rational standards for human judgment in real-world tasks. Proposing heuristics as descriptive models to account for deviations from statistical norms only clouds the issue, as there is no way to tell whether an apparent deviation is a poor heuristic approximation to a presumed statistical norm, or a good approximation to some more adaptive approach.

We will propose that a *causal Bayesian framework* provides this more adaptive approach, and that it offers both a better normative standard than purely statistical methods and a better descriptive model than heuristic accounts. Like classical statistical norms, the framework we propose is Bayesian, but rather than making inferences from purely statistical data, inferences in our framework are made with respect to a causal model, and are subject to the constraints of causal domain knowledge. Our causal Bayesian framework is more adaptive than previous proposals, as it explains how rational judgments can be made with the relatively limited statistical data that is typically available in real-world environments. This approach also represents a better descriptive model than purely statistical norms or heuristics, which do not

emphasize, or even have room to accommodate, the kinds of causal knowledge that seem to underlie much of people's real-world judgment.

We study the role of causal knowledge in judgment by focusing on how it modulates the classic phenomenon of "base-rate neglect". Early studies indicated that people were more likely to neglect base rates that lack "causal relevance" (Ajzen, 1977; Tversky & Kahneman, 1980), although the notion of causal relevance was never well defined. Bar-Hillel (1980) argued that the salience of the base rate determined whether people would use this information, and that causal relevance was just one form of salience. Contrary to the conclusions of the heuristics & biases literature, we argue that for many well-known stimuli, the features of the base rate are not what lead people to exhibit apparent "base-rate neglect". We offer as evidence four experiments in which the description of the base rate is identical across two conditions, but people neglect the base rate in one condition and use it appropriately in the second condition. We further argue that people in these experiments may not actually be misusing base rate statistics; we show how cases of "base-rate neglect" may be re-interpreted as cases in which the prescriptions of classical Bayesian norms are non-normative by the standards of causal Bayesian inference. Our experiments will show that when these prescriptive norms agree, people often use the given statistics normatively (by both standards), but when they disagree, people's judgments more often adhere to the causal Bayesian standard than the classical Bayesian standard. Furthermore, when the problem makes clear which causal model should be used and how given statistics should be incorporated into that model, we find that people rarely neglect base rates.

As indicated above, we are not the first to propose that causal knowledge plays a role in base-rate neglect. Researchers in the heuristics and biases tradition investigated how causal factors influenced the phenomenology of base-rate neglect, but they offered no precise or

generalizable models for how causal knowledge and probabilistic judgment interact, and they did not explore the rational role of causal reasoning in judgment under uncertainty. Ajzen (1977) proposed that a “causality heuristic” leads to neglect of information that has no apparent causal explanation. Following Ajzen (1977), Tversky and Kahneman (1980) proposed that “evidence that fits into a causal schema is utilized, whereas equally informative evidence which is not given a causal interpretation is neglected” (Tversky & Kahneman, 1980, p. 65). However, neither explained how a focus on causal factors could lead to successful judgments in the real world. They did not attempt to explain why people would have such a heuristic or why it should work the way that it does. On the contrary, the heuristics and biases tradition did not appear to treat attention to causal structure as rational or adaptive. The use of causal schemas was instead viewed as an intuitive, fuzzy form of reasoning that, to our detriment, tends to take precedence over normative statistical reasoning when given the chance. In contrast to Tversky and Kahneman’s (1980) undefined “causal schemas”, our proposal for inference over causal models based on Bayesian networks provides a well-defined, rational, and adaptive method for judgment under uncertainty, which can succeed in real-world tasks where noncausal statistical methods fail to apply. We will argue that people’s judgments can in fact be both causally constrained *and* rational – and rational precisely because of how they exploit causal knowledge.

#### A causal Bayesian framework for judgment under uncertainty

Causal reasoning enables one to combine available statistics with knowledge of causal relationships, resulting in more reliable judgments, with less data required than purely statistical methods. It is becoming clear from research in artificial intelligence (Pearl, 2000), associative learning (Cheng, 1997; Glymour, 2001; Gopnik & Glymour, 2002; Gopnik & Sobel, 2000; Waldmann, 1996), and categorization (Ahn, 1999; Rehder, 2003) that causal reasoning methods

are often better suited than purely statistical methods for inference in real-world environments. Causal Bayesian networks have been proposed as tools for understanding how people intuitively learn and reason about causal systems (e.g. Glymour & Cheng, 1998; Gopnik, et al., 2004; Griffiths & Tenenbaum, 2005; Sloman & Lagnado, in press; Steyvers, Tenenbaum, Wagenmakers & Blum, 2003; Tenenbaum & Griffiths, 2001, 2003; Waldmann, 2001), but their implications for more general phenomena of judgment under uncertainty have not been systematically explored. We see the present paper as a first attempt in this direction, with a focus on explaining base-rate neglect, one of the best-known phenomena of modern research on probabilistic judgment.

This section outlines the theoretical background for our work. A full and formal treatment of causal Bayesian networks is beyond the scope of this paper, so we begin by summarizing the main aspects of the causal Bayesian framework that our experiments build on. We then argue that this framework represents a better normative standard for judgment under uncertainty in real-world environments than the purely statistical frameworks that have motivated previous research on human judgment. Finally, we will illustrate how causal Bayesian reasoning implies that a given piece of statistical information, such as a base rate, may be used differently depending on the reasoner's causal model. In particular, we describe two ways in which the causal Bayesian framework may predict cases of apparent base-rate neglect: a given statistic may not always fit into a person's causal model, or it may fit in such a way that the structure of the causal model dictates that it is not relevant to a certain judgment at hand.

We will study human judgment as an approximation to the following ideal analysis. We assume that a causal mental model can be represented by a Bayesian network (Pearl, 2000), a directed graphical probabilistic model in which nodes correspond to variables and edges

correspond to direct causal influences. A set of parameters associated with each node defines a conditional probability distribution for the corresponding variable, conditioned on the values of its parents in the causal graph (i.e., its direct causes). Loosely speaking, the edge structure of the graph specifies what causes what, while the parameters specify how effects depend probabilistically on their causes. The product of the conditional distributions associated with each node defines a full joint probability distribution over all variables in the system. Any probabilistic judgment of interest can be computed by manipulating this joint distribution in accordance with Bayes' rule.

When people are confronted with a judgment task, they face three distinct aspects of judgment: (1) constructing a causal model, (2) setting the model's parameters, and (3) inferring probabilities of target variables via Bayesian inference over the model. We will illustrate this process using a scenario we created for Experiment 1, a version of the classic problem in which participants are asked to judge the probability that a woman receiving a positive mammogram actually has breast cancer, which we call the "benign cyst" scenario. The text of the scenario is shown in Figure 1, along with the three stages of judgment.

The first step is to construct a causal model (see Figure 1a) relating the variables described in the task. In this case, information is given in the task description that benign cysts (the variable *cyst*) can cause positive mammograms (the variable *+M*). Often, however, people rely on prior knowledge of which variables cause which others. For example, most participants already know that breast cancer (the variable *cancer*) causes positive mammograms (the variable *+M*). The second step is to set the values of the parameters characterizing the relationships between cause and effect (see Figure 1b). In this case precise numerical values for some parameters can be determined from the statistical information provided in the judgment task (e.g.,

the base rate of breast cancer is 2% and the base rate of benign cysts is approximately 6%, neglecting the very low probability that a woman has both breast cancer and benign cysts). Background knowledge may be necessary to supply values for other parameters, such as how the effect depends on its causes. In this case, one might assume that positive mammograms do not occur unless caused, that the causes act independently to generate positive mammograms, and that both of the given causes are fairly strong. These assumptions can be captured using a noisy-or parameterization (Pearl, 1988), a simple model to characterize independent generative causes that is equivalent to the parameterizations used in Cheng's (1997) power PC theory or Rehder's (2003) causal models for categorization. For simplicity, in Figure 1 we assume that either cause when present produces the effect with probability near 1. Once the causal structure of the model has been determined and parameters have been specified, inference can be performed on this model to make judgments about unknown variables (see Figure 1c). Bayesian reasoning prescribes the correct answer for any judgment about the state of one variable given knowledge about the states of other variables. (e.g., given knowledge that a woman received a positive mammogram, the probability that she has cancer as opposed to a benign cyst is approximately 25%).

Expert systems based on Bayesian networks (Pearl, 1988) have traditionally been built and used in just this three-stage fashion: an expert provides an initial qualitative causal model, objectively measured statistics determine the model's parameters, and inference over the resulting Bayesian network model automates the expert's judgments about unobserved events in novel cases. This top-down, knowledge-based view of how causal models are constructed is somewhat different from much recent work on causal learning in psychology, which emphasizes more bottom-up mechanisms of statistical induction from data (e.g., Glymour, 2001). As we will

argue, however, the prior causal knowledge that people bring to a judgment task is essential for explaining how those judgments can be successful in the real world, as well as for determining when and how certain statistics given in a task (such as base rates) will affect people's judgments.

*Causal Bayesian inference as a new normative standard*

This causal Bayesian framework may provide a more reasonable normative standard for human judgment than classical, purely statistical norms that do not depend on a causal analysis. Our argument follows in the tradition of recent rational analyses of cognition (Anderson, 1990; Oaksford & Chater, 1999). Causal Bayesian reasoning is often more ecologically adaptive, because it can leverage available causal knowledge to make appropriate judgments even when there is not sufficient statistical data available to make rational inferences under non-causal Bayesian norms. The structure of the causal model typically reduces the number of numerical parameters that must be set in order to perform Bayesian inference. In general, for a system of  $N$  variables, standard Bayesian inference using the full joint distribution requires specifying  $2^N - 1$  numbers, while a causally structured model could involve considerably fewer parameters. For instance, in Figure 1, only four parameters are needed to specify the joint distribution among three variables, which would require seven numbers if we were using conventional Bayesian reasoning. The simplifications come from several pieces of qualitative causal knowledge: that there is no direct causal connection between having breast cancer and having benign cysts, that breast cancer and benign cysts act to produce positive mammograms through independent mechanisms (and hence each cause requires only one parameter to describe its influence on the effect of a positive mammogram), and that there are no other causes of positive mammograms.

In more complex, real-world inference situations there are often many relevant variables, and as the number of variables increases this difference becomes more dramatic. We cannot generally expect to have sufficient data available to determine the full joint distribution over all these variables in a purely statistical, non-causal fashion, with the number of degrees of freedom in this distribution increasing exponentially with the number of variables. Causal Bayesian inference provides a way to go beyond the available data, by using causal domain knowledge to fill in the gaps where statistical data is lacking.

*Causal Bayesian inference as a descriptive model*

The Causal Bayesian framework can be used to explain human judgment in the “base rate neglect” literature and in our experiments. Because judgments are made using causal models, rather than statistics alone, experimenter-provided statistics should be used differently depending on the causal structure one believes to be underlying them. Two cases are of particular importance to base rate neglect, and will be explored in our experiments.

First, statistics can often map clearly onto one causal model, but not another. For example, suppose we are provided with statistical data indicating that the risk of breast cancer ( $C$ ) is associated with two lifestyle factors, being childless ( $L$ ) and having high stress levels ( $S$ ), but we do not have the full joint distribution among the three variables. Suppose further that we know that stress causes cancer, but we do not know how being childless and having cancer are causally related. Three causal models are consistent with this knowledge (see Figure 2), but they have different parameters, which means a given statistic may fit clearly into one model but not into another. Suppose for instance we are given the statistic for the probability of high stress levels given that one is childless, (e.g.,  $P(S | L) = 0.75$ ). For Figure 2b, the given statistic corresponds directly to a model parameter, thus it can be directly assigned to that parameter. For Figure 2c,

there is no model parameter corresponding to  $P(S|L)$ , but there is a parameter corresponding to its inverse,  $P(L|S)$ , hence one can assign the formula  $P(S|L)P(L)/P(S)$  to the parameter for  $P(L|S)$ . For Figure 2a,  $P(S|L)$  does not correspond directly to a parameter of the model or its inverse, which means there is no single prescription for how such a statistic will influence future judgments from this model. In Experiments 1, 2, and 3, we test the hypothesis that statistics that map clearly onto parameters of the causal model are more likely to be used appropriately, while statistics that do not have corresponding parameters in the model are more likely to be used incorrectly or ignored.

Second, even when provided statistics can be clearly mapped to parameters, the causal Bayesian framework prescribes different ways of using those statistics in making judgments depending on the causal structure. Suppose we are told that a particular woman is childless, and we are asked to judge the likelihood of her having cancer. If being childless causes breast cancer (Figure 2a), then the risk of cancer in a childless woman is increased, regardless of whether or not she has high stress levels (assuming for simplicity that these factors do not interact). However, it could be the case that being childless causes women to develop high stress levels (Figure 2b), but does not directly cause cancer. In this case, the risk of cancer in a childless woman is still increased, but we can ignore the fact that she is childless if we know her level of stress. Finally, it might be the case that having high stress levels causes women not to have children (Figure 2c). In this case, we should again ignore the fact that a woman is childless if we already know the woman's level of stress. These principles of causal structure are intuitively sound, but the notion that statistical data should be used differently for different causal structures is beyond the scope of classical statistical norms. The causal Bayesian standard is able to make inferences where previous standards could not, prescribing the appropriate use of limited data by

making use of the conditional independence relations determined by causal structure. Experiment 4 tests the applicability of this principle to human judgment. We test the hypothesis that people will ignore a given piece of statistical data in a case where it is rational to do so given the causal structure of the task, but they will use that same statistic if the causal structure is slightly modified to suggest that it is relevant to the judgment at hand.

### Experiments

Psychological studies of judgment under uncertainty typically provide people with statistical data and then ask them to make judgments using the provided statistics, but if several different causal models are possible, this information may not be sufficient for the causal Bayesian framework to prescribe a unique correct answer. The normatively correct inference depends both on the particular causal model used and on how the statistics are assigned to the parameters of the model. Therefore it is only possible to prescribe a single correct answer using the causal Bayesian framework if (1) the model structure is known, (2) the provided statistics map unambiguously onto model parameters, and (3) no free parameters remain after assigning the provided statistics to the problem. The issue of how provided statistics are used to update the parameters of a model, and how they are then used in subsequent inferences, plays a central role in our experiments. In each of the following four experiments we test the extent to which people's judgments conform to the prescriptions of causal Bayesian inference by providing a scenario in which the statistics clearly map onto the parameters of an unambiguous causal model. In Experiments 1-3 we compare these judgments to those on an equivalent scenario from the base-rate neglect literature in which the statistics do not map clearly onto parameters of the causal model. In Experiment 4, we compare these judgments to those on an equivalent scenario

with a different causal structure, in which the base rate statistic is rendered irrelevant to the judgments.

In each experiment, the formal statistical structures of the two scenarios were always identical from the point of view of the classical Bayesian norm, thus the judgment prescribed by this norm is the same for the two scenarios. Furthermore, all other factors previously identified as playing a role in base-rate neglect (such as salience or causal relevance of the base rate) were held constant, thus the heuristics & biases view would predict that people exhibit identical levels of base rate neglect in the two scenarios. Crucially, however, the two scenarios always differ in their causal structure, such that the correct answers prescribed by our new causal Bayesian norm differ across scenarios. Thus, only the causal Bayesian framework predicts that people's judgments will differ between the two scenarios. In addition, only the causal Bayesian framework predicts that people will exhibit less base-rate neglect on the scenario with a clear parameter mapping and a causal structure that requires that the base rate be used.

### Experiment 1

In the original mammogram problem (Eddy, 1982; Gigerenzer & Hoffrage, 1995) the base rate of cancer in the population often appears to be neglected when people judge the likelihood that a woman who receives a positive mammogram has cancer. Figure 3(a-c) depict the three phases of inference for the mammogram scenario. A causal model of this scenario constructed from common knowledge should include *cancer* as a cause of positive mammograms (+*M*), as depicted Figure 3a. In this model, the variable *cancer* has no parents, therefore the CPT for *cancer* contains just one parameter:  $P(\textit{cancer})$ , which directly corresponds to the base rate provided in the problem. Because there is only one way to assign the base rate to this model parameter, the base rate should influence judgments by causal Bayesian standards.

In this experiment, we demonstrate empirically that people do not neglect this base rate in a newly developed scenario that differs only in causal structure, and we argue that the real difficulty people have in the classic version is with the false-positive statistic: the probability of a positive mammogram in a woman who does not have cancer,  $P(+M | \neg cancer) = 9.6\%$ . On classic versions of this task, we hypothesize that people may adopt a causal model in which the false-positive statistic does not correspond to a model parameter. If people assume a noisy-or parameterization, as we used in constructing the causal model in Figure 1, the model will have a parameter for the  $P(+M | cancer)$  but not for  $P(+M | \neg cancer)$ ; this reflects the intuition that the absence of a cause has no power to produce an effect (see Figure 3b). Although it would be possible to accommodate the false-positive statistic within a noisy-or parameterization by hypothesizing an additional cause of positive mammograms, and interpreting this statistic as the causal power of that alternative cause, this represents several steps of hypothetical reasoning that might not occur to many people.<sup>2</sup> Many participants may realize that the false positive statistic is somehow relevant, and if they cannot fit it into their model, they may look for some simple way to use it to adjust their judgment. For example, subtracting the false-positive rate from the true positive rate would be one such strategy, consistent with typical responses classified as “base-rate neglect”.

To test our hypothesis about when people can use base rates properly in causal inference, we developed a new scenario in which the causal model is clear and all the statistics clearly map onto parameters of the model. We clarified the causal structure by providing an explicit alternative cause for positive mammograms in women who do not have cancer: benign cysts (see Figure 1). We replaced the false-positive rate in the original problem with the base rate of dense but harmless cysts, and described the mechanism by which these cysts generate positive

mammograms. This new statistic, the base rate of benign cysts in the population, directly maps onto the parameter for  $P(\text{cyst})$  in the model (see Figure 1b).

### *Method*

*Participants.* The participants in this experiment were 60 MIT undergraduate and graduate students. They were recruited randomly in a main corridor on campus and given token compensation.

*Materials.* Participants were randomly assigned to receive one of two paper and pen versions of Gigerenzer and Hoffrage's (1995) probabilistic mammogram question (adapted from Eddy, 1982) in a between-subjects design. The *false positive scenario* was similar to the original question, while the *benign cyst scenario* gave the base rate of benign cysts in the population rather than the rate of false positives. We chose not to include the true-positive rate,  $P(+M | \text{cancer}) = 0.8$ , but instead stated, "most women with breast cancer will receive a positive mammogram." This was done to encourage participants to provide answers based on their intuition rather than memorized mathematical formulas. Both scenarios required the exact same mathematical formula to calculate the answer, so there was no difference in arithmetic difficulty. The exact wording of the scenarios is shown in Figure 3, along with the three phases of judgment prescribed by causal Bayesian inference. We also asked participants to rate the believability of the statistics given, to ensure that the benign cyst statistic and the false positive statistic were equally believable (see below).

### *Believability ratings*

The problem on the previous page gave several statistics about breast cancer and mammogram tests. To what extent do you think the given statistics accurately reflect the real-world probabilities of these events?

Answer on a scale of 1 to 7 (circle one):

1	2	3	4	5	6	7
Very inaccurate			Not sure			Very accurate

### Results

The results show significantly improved performance on the benign cyst scenario. The correct answer to both scenarios is (by the standards of classical Bayesian inference)

approximately  $\frac{2\%}{2\% + 98\% \times 6\%} \approx \frac{2\%}{2\% + 6\%} = 25\%$ . We classified as *correct* answers of exactly

25% (with rounding), and classified as *base-rate neglect* any answer over 80% (however there were no responses between 80% and 90%). We also found several answers of *odds form* (answers of 2/6 or 33%) and *base rate overuse* (answers of 2%). All remaining responses were classified as *other* (answers of *odds form* and *base rate overuse* are shown in Figure 4, but were collapsed with *other* for the statistical analysis). Responses consistent with *base-rate neglect* were significantly lower and *correct* answer rates were significantly higher on the benign cyst scenario compared to the false positive scenario ( $\chi^2(2) = 6.29$ ,  $p < .05$ , see Figure 4).

Since people likely assume error rates to be low in medical tests, one explanation for our results is that people find the benign cyst statistic more consistent with their prior beliefs than the false positive statistic. However, the results of our believability question indicate no significant difference in believability ratings between the two scenarios (4.03 average rating for the benign cyst scenario vs. 4.37 average rating for the false positive scenario). Thus, it is unlikely that believability can serve as an explanation for the difference in performance.

### *Discussion*

These results support our hypothesis that people are adept at making rational judgments from statistics that unambiguously map onto parameters of a clear causal model. We found that the modal response to the benign cyst scenario was the *correct* answer, with only 2 out of 30 responses consistent with *base-rate neglect*. We also replicated previous findings (Gigerenzer & Hoffrage, 1995) that the modal response to the original scenario was consistent with *base-rate neglect*, demonstrating that our participants were not merely more statistically adept than the populations of previous studies. In a previous study, Krynski & Tenenbaum (2003) obtained similar and even more significant results on essentially the same question with 157 participants containing a mix of airport passengers and MIT students.

The heuristics and biases view cannot account for these results, which, by classical Bayesian standards, show increased *correct* responses and fewer responses consistent with *base-rate neglect* on our newly developed benign cyst scenario. While causality has been implicated as a factor in *base-rate neglect*, the focus has been only on the base rate statistic itself. Tversky and Kahneman (1980) state that “base-rate data that are given a causal interpretation affect judgments, while base-rates that do not fit into a causal schema are dominated by causally relevant data” (p. 50). Since the base rate of cancer itself is presented identically in the two scenarios, the heuristics and biases view cannot explain why it is more often used properly in the benign cyst scenario while so often “neglected” in the false positive scenario.

### Experiment 2

In Experiment 1, we demonstrated that people often make accurate judgments about the uncertainty of a given cause being responsible for an observed effect. However, the mechanisms involved were described in essentially deterministic terms. In Experiment 2, we introduce a

second source of uncertainty: probabilistic mechanisms. We again created a version of the mammogram problem (Eddy, 1982; Gigerenzer & Hoffrage, 1995), but this time we included the true positive rate,  $P(+M | cancer)$ , as well as the propensity of a benign cyst to cause a positive mammogram,  $P(+M | benign\ cyst)$ . This enabled us to test whether people reason appropriately about the uncertainty introduced by probabilistic mechanisms. It also made the difficulty of the questions more similar to prior research, and provided a more rigorous test of people's judgment capabilities.

We also attempted to eliminate some potential confounds with the previous experiment. In Experiment 1, the salience and descriptive detail of the causal mechanism may have enabled people to pay more attention or think more clearly about the benign cyst scenario, which could account for their improved performance by classical Bayesian standards. For Experiment 2, both questions were written with only dry statistical information; instead of the description of the causal mechanism by which a benign cyst can lead to a positive mammogram, we provided only statistical data concerning the rate of benign cysts in the population and the likelihood of them causing a positive mammogram.

### *Method*

*Participants.* The participants in this experiment were 59 MIT undergraduates. They were recruited during a study break in an undergraduate dormitory, and were given token compensation.

*Materials.* We again randomly assigned participants to one of two scenarios based on Gigerenzer and Hoffrage's (1995) probabilistic mammogram question (adapted from Eddy, 1982). The *false positive scenario* was similar to the original question, while the *benign cyst scenario* gave the base rate of benign cysts in the population, as well as a likelihood of a positive

mammogram given a benign cyst, rather than the rate of false positives. Under the classical Bayesian norm, the two calculations were equally difficult, ensuring that any improvement on the benign cyst scenario over the false positive scenario would not be due to an easier computation. The correct Bayesian calculation for the false positive scenario was:

$$P(H | D) = \frac{P(H) * P(D | H)}{P(H) * P(D | H) + P(\neg H) * P(D | \neg H)}$$

while the (approximately) correct Bayesian calculation for the benign cyst scenario was:

$$P(H | D) = \frac{P(H) * P(D | H)}{P(H) * P(D | H) + P(A) * P(D | A)} \quad (\text{where } A \text{ is the alternate cause, benign cysts})$$

#### *False Positive Scenario*

Doctors often encourage women at age 50 to participate in a routine mammography screening for breast cancer. From past statistics, the following is known:

1% of the women had breast cancer at the time of the screening

Of those with breast cancer, 80% received a positive result on the mammogram

Of those without breast cancer, 15% received a positive result on the mammogram

All others received a negative result

Suppose a woman gets a positive result during a routine mammogram screening.

Without knowing any other symptoms, what are the chances she has breast cancer?

#### *Benign Cyst Scenario*

Doctors often encourage women at age 50 to participate in a routine mammography screening for breast cancer. From past statistics, the following is known:

1% of the women had breast cancer at the time of the screening

Of those with breast cancer, 80% received a positive result on the mammogram

30% of the women had a benign cyst at the time of the screening

Of those with a benign cyst, 50% received a positive result on the mammogram

All others received a negative result

Suppose a woman gets a positive result during a routine mammogram screening.

Without knowing any other symptoms, what are the chances she has breast cancer?

### *Results*

The correct answer to the both scenarios is  $\frac{1\% \times 80\%}{1\% \times 80\% + 15\%} \approx 5.1\%$  (approximately). We

classified as *correct* any rounded version of the exact answer (e.g., 5%, 5.1%, 5.06%, etc), and we again classified as *base-rate neglect* any answer over  $P(D|H) - P(D|\neg H)$ , the lower bound of typical base-rate neglect answers (65% in this problem). All other answers were classified as *neither*, and were distributed across the spectrum with no discernable pattern. Our results show that *base-rate neglect* was significantly lower and *correct* answer rates were significantly higher on the benign cyst scenario (Fisher's Exact Test<sup>3</sup>,  $p < .05$ , see Figure 5). The size of the effect was also considerable, with *correct* responses more than doubled (11 of 30 on the benign cyst scenario vs. 4 of 29 on the false positive scenario), and responses consistent with *base-rate neglect* nearly eliminated (1 of 30 on the benign cyst scenario vs. 7 of 29 on the false positive scenario).

### *Discussion*

The results of Experiment 2 again support our hypothesis that people generally make judgments consistent with the causal Bayesian framework. The benign cyst scenario in this case provided two sources of uncertainty: (1) multiple causes of an observed effect, and (2) probabilistic mechanisms by which those causes produce the effect. Although the arithmetic complexity involved prevented many participants from providing exactly correct responses,

performance was relatively good overall, suggesting that people are often able to interpret and reason appropriately about statistics describing multiple probabilistic mechanisms.

Experiment 2 also addressed the potential confound with Experiment 1 that the benign cyst scenario was more descriptive of the mammogram mechanism than the false positive scenario, and hence potentially more salient. In Experiment 2, we did not describe the mechanism, and neither did we explicitly state that benign cysts were an alternative cause for positive mammograms. Merely by structuring the statistics in terms of an alternate cause, as opposed to a rate of false positives, we were able improve performance (according to classical Bayesian norms). This suggests that if the existence of a mechanism seems plausible, such as a mechanism by which benign cysts can generate positive mammograms, one need not understand the mechanism entirely to reason about it appropriately.

### Experiment 3

Experiments 1 and 2 demonstrated that statistics that can be unambiguously assigned to causal model parameters are often used appropriately. This finding bears some resemblance to previous appeals to “causal relevance” in explaining base-rate neglect (Ajzen, 1977; Tversky & Kahneman, 1980), but our causal Bayesian account is importantly different. Under previous “causal relevance” accounts the prescribed method of judgment is the use of Bayes’ rule and base rates are neglected if they do not fit into a causal schema. In contrast, the prescribed method of judgment in our framework is Bayesian inference over causal models, which means sometimes base rates can appear to be neglected even if they do fit into the model because they may not be required in the final judgment. This is what we propose happens when people seem to neglect the base rate in the “cab problem” (Kahneman & Tversky, 1972).

The “cab problem” was one of the earliest problems found to elicit base-rate neglect. In this problem, participants were told that a witness identified the color of a cab in hit-and-run accident, claiming it was blue, although only 15% of the cabs in the city were blue and the remaining 85% were green. In subsequent tests of the witness’ vision, the witness mistakenly identified 20% of the green cabs as blue and 20% of the blue cabs as green. Participants were then asked to judge the likelihood that the cab was really blue. Participants famously tended to ignore the base rate of 15% blue cabs in the city, providing a modal response of 80% chance that the cab was really blue. In a follow-up study, Tversky & Kahneman (1980) obtained improved performance using a “causal” base rate: now 50% of the cabs in the city were blue and 50% were green, but only 15% of the cabs involved in accidents were blue and 85% were green.

Tversky and Kahneman (1980) argued that the population base rate in the original problem is neglected because it does not fit into a causal schema; i.e., nothing causes there to be more green cabs. In contrast, they argued, the base rate in the causal cab problem fits into a causal schema: the higher accident rate of green cabs might be caused by Green cab drivers being more reckless. One major difficulty with this proposal is that it is unclear what it means for a statistic to fit into a causal schema. In particular, Tversky and Kahneman make the assumption that population base rates do not fit into causal schemas. However, causal Bayesian networks require base rates for uncaused variables, thus the provided base rates should fit equally well into the causal models for both scenarios. Furthermore, intuitive causal judgment is clearly sensitive to base rates; when judging whether someone’s cough is more likely to be caused by a common cold or by lung cancer, the base rate of these illnesses in the population is obviously essential. Further casting doubt on the ability of causal schemas to explain base rate neglect, Bar-Hillel

(1980) showed that even population base rates can elicit good performance in the cab problem (using a third variant, the “intercom problem”).

In this experiment, we demonstrate that the base rate in the original cab problem, previously labeled “non-causal”, can be easily interpretable and used properly. Our explanation for the previous findings of base-rate neglect starts with the observation that the causal structure of this problem is more complex than anticipated by previous researchers. It is common knowledge that perceptual discriminations are strongly influenced not only by the intrinsic discriminability of objects in the world but also by prior knowledge, such as stereotypes or more mundane expectations. For instance, if you feel a rough oblong object in a fruit bowl you might guess it is a pear, whereas if you feel the same object in a vegetable drawer you might guess it is a potato. Since prior expectations influence people’s judgments, unexpected judgments can be as accurate as expected ones. Someone who thinks he found a pear in a vegetable drawer probably inspected it more closely than someone who thinks he found a potato in that drawer. More formally, signal detection theory captures the contributions of both discriminability ( $d'$ ) and prior expectations (through criterion shifts of the area under the occurrence distributions). Using a causal model to capture this process, there are three causes of the witness’ judgment (see Figure 6b). In this case, the given base rate statistics can be applied to the parameters for the *witness’ expectations*, while the given statistics for confusing blue and green can be applied to the parameters for *discriminability*. Because the model is so complex, these statistics end up being used differently than if they were inserted into Bayes’ rule according to the classical Bayesian norm. According to Birnbaum (1983), if the witness’ judgment can be characterized by signal detection theory then ignoring the base rate of cabs may actually be justified. Birnbaum’s (1983) analysis suggests that for an optimizing witness whose expectations match the true base rate, the

probability of the witness being correct remains at approximately 80% across a large range of base rates. This means ignoring the base rate may actually be appropriate, because the witness has essentially already taken it into account.

Although this account may be plausible, it is essentially speculative because we do not know what causal model people have in mind when solving these problems. It could be that people really have a signal detection model in mind, and their judgments are accurate by that model. Alternatively, people could merely have the intuition that reliability in perception should not vary across base rates. Only by clarifying the causal structure can we make clear predictions using the causal Bayesian framework. In our new *faded paint scenario*, 20% of green cabs have faded paint, making them appear blue, and 20% of blue cabs have faded paint, making them appear green, while the witness accurately reports the apparent color of the cab. This results in an identical computation (under classical Bayesian inference) to the original cab problem, but since it is clear how the model should be structured and how the given statistic should be assigned to its parameters, a correct solution can be prescribed using the causal Bayesian framework. This correct solution requires use of the base rate, and is the same as the solution prescribed by classical Bayesian inference. The three phases of causal Bayesian inference for this scenario are depicted in Figure 7(a-c). Like the previous experiments, we tested people's performance on our new scenario while using the original *perceptual error scenario* as a control.

### *Method*

*Participants.* The 47 participants in this experiment were MIT undergraduate and graduate students (majors were not recorded, but were likely randomly distributed). They were approached in a student center, and given token compensation.

*Materials.* Participants were randomly assigned to receive one of two variants of Kahneman and Tversky's cab problem in a between-subjects design. The *perceptual error scenario*, modeled after the original cab problem, attributed the witness' mistakes to perceptual error. The *faded paint scenario* attributed the witness' mistakes to faded paint. Both questions require the exact same calculation, so they are equally difficult to solve with classical Bayesian methods. The questions follow:

### *Results*

We classified as *correct* only answers that matched (with rounding) the normative solution by classical Bayesian standards:  $\frac{15\% \times 80\%}{15\% \times 80\% + 85\% \times 20\%} = \frac{12\%}{29\%} = 41\%$ . Consistent with prior studies, we classified as *base-rate neglect* any response of 80% or above. All other answers were classified as *neither*, and were distributed across the spectrum with no discernable pattern. The results show a significant reduction in *base-rate neglect* and a significant increase in *correct* responses for the faded paint scenario ( $\chi^2(2) = 11.55, p < .005$ ) (see Figure 8). The effect was very large, with *correct* responses dramatically improved (11 of 24 on the faded paint scenario vs. 2 of 23 on the perceptual error scenario), and responses consistent with *base-rate neglect* nearly eliminated (2 of 24 on the faded paint scenario vs. 10 of 23 on the perceptual error scenario).

### *Discussion*

The results show generally strong performance on the faded paint scenario, with nearly half the participants giving the exact correct solution on a mathematically difficult task. Using the original scenario as a control, we found that our participants are not merely more statistically adept than other populations, as they exhibited the classic pattern of responses consistent with base-rate neglect on the original scenario. These results are not predicted by the heuristics and

biases view, which predicts that people use base rates that seem causally relevant, but neglect those that do not; in this case, the base rates are identical, yet people perform much better on the faded paint scenario.

Like in the previous experiments, we see important implications not just for the role of causality in judgments tasks, but also for the role of statistics in judgments from causal models. The cab problem is controversial in part because perception is a much more complicated process than can be summed up in two conditional probabilities. People's neglect of the base rate suggests that people may have a special model for perception in which accuracy remains relatively constant across base rates, but varies with the confusability of the items to be perceived, as would be expected if people are optimal signal detectors. This perceptual model may be one of many such models that people have for a variety of specific domains that enable them to make reasonably accurate qualitative judgments about complex mechanisms.

#### Experiment 4

In this experiment, we demonstrate more directly that causal structure should influence whether base rates will be used, and that the classical Bayesian framework, which ignores causal structure, can sometimes fail to provide an appropriate standard for judgment. The previous three experiments all compared scenarios that differed considerably in their descriptions, introducing possible imbalances in salience or plausibility that could be implicated in our results. In this experiment, we adopt stimuli that are otherwise identical to each other except for a manipulation of causal structure and requested judgment. All the stimuli were newly developed, modeled after social judgment tasks such as Kahneman and Tversky's (1973) famous lawyer-engineer problem, which asked participants to predict a man's career from his personality. In the lawyer-engineer problem, participants were told in one condition that personality tests were administered to 30

engineers and 70 lawyers in one condition, while in another condition the tests were administered to 70 engineers and 30 lawyers. Participants were then presented with several personality profiles and asked to judge the likelihood that each was a lawyer or an engineer. Kahneman and Tversky's (1973) findings were that although people gave slightly different responses in the two conditions, they were not different enough. According to classical Bayesian principles, the ratio of base rates of lawyers to engineers should dramatically influence judgments of career given personality.<sup>4</sup>

In this experiment we investigated a new problem modeled after the lawyer-engineer problem. The cover story described a CIA special operations team selected its members according to certain criteria. In the first scenario, part of the mission required female agents, while in the second scenario, part of the mission required short agents. Causal models for these two scenarios are depicted in Figure 10 and Figure 11. Because women are shorter than men, the agents in the first scenario were mostly 5'7" or under. Because it was easier to find short women than short men, the agents in the second scenario were mostly women.

For each scenario, two different tasks were posed (each participant received only one task). The first task was to judge the probability that a male agent randomly selected from the team is shorter than average. According to the causal Bayesian framework, in the judgment of height from gender, one should ignore the base rate of heights in the first scenario, but not in the second scenario. Intuitively, this makes sense: a male agent on the team *selected by gender* is no more likely to be short than any other male agent, even if the base rate of people over 5'7" is 30%. However, a male member on the team *selected by height* is likely to be shorter than average, thus the difference in base rate of heights matters. This distinction between the scenarios derives from their causal structures, depicted in Figure 10 and Figure 11. Given the causal model for the

*selected by gender* scenario (see Figure 10(a)), learning the gender of an agent renders the fact that the agent is on the team irrelevant to judging the agent's height, because the path from *on CIA team* to *height* is blocked by *gender* (technically, the causal markov condition of Bayesian networks specifies that *on CIA team* is conditionally independent of *height* given *gender*). Because it is irrelevant that the agent is on the team, the base rate of heights on the team is also irrelevant. In contrast, given the causal model for the *selected by height* scenario (see Figure 11 (a)), after learning the gender of an agent the fact that the agent is on the team remains relevant to judging the agent's height because the path from *on CIA team* to *height* is not blocked by *gender*.

The second task was to judge whether, compared to other CIA teams, a randomly selected agent over 5'7" on this team is more likely to be male. While the first task judged height from gender, this task judged gender from height. By reversing which variable was observed and which was judged, we also reversed which scenario required ignoring the base rate; in the first task the *selected by gender* scenario required ignoring the base rate, while in this task the *selected by height* scenario required ignoring the base rate. Intuitively, because the team *selected by gender* has more females, more of the tall agents are female than on other teams, thus the base rate of gender matters. Conversely, a tall person on the team *selected by height* is no more likely to be male than on any other team, assuming that gender does not influence selection for the team. Formally, in the *selected by height* scenario the path from *gender* to *on CIA team* is blocked by *height*, therefore the fact that the tall person is on the team is irrelevant for judging the person's gender, and the base rate of heights on the team should be ignored. In contrast, in the *selected by gender* scenario the path from *gender* to *on CIA team* is not blocked by height,

therefore the fact that the tall person is on the team is relevant to judging the person's gender, and the base rate of heights on the team should be used.

With Experiment 4, we used the gender-height scenarios to test whether people use or ignore the base rate appropriately according to the causal Bayesian framework. This framework prescribes that people should ignore the base rate when the observed variable blocks the path between the variable to be judged and the variable representing being on the team, and should use the base rate when the variable to be judged falls between the observed variable and the variable representing being on the team.

### *Method*

*Participants.* The 152 participants in this experiment were MIT undergraduates and graduate students (majors were not recorded, but were likely randomly distributed). They were approached in a main corridor on campus, and were given token compensation.

*Materials:* Participants were assigned randomly to one of four conditions, as we crossed two factors: whether the team members were selected by gender or height (the causal structure), and whether the judgment was to infer gender from height or height from gender (the judgment). The cover story changed depending on the causal structure, while the question changed depending on the judgment. The materials are depicted in Figure 10 and Figure 11, along with a depiction of the three phases of judgment prescribed by causal Bayesian inference.

### *Results*

The results are depicted in Figure 12. The modal response in all conditions matched the predictions of our causal Bayesian hypothesis. For judgments of height from gender, there were three possible responses: *shorter*, *taller*, or *no different*. A significant difference was found in responses levels between the two scenarios ( $\chi^2(2)=9.7$ ,  $p<.001$ ). Responses of *no different*

(indicating base rate ignorance) were higher for the *selected by gender* causal structure than the *selected by height* causal structure, which is consistent with the prescription of the causal Bayesian norm that when selected by gender, knowing the agent's *gender* renders *height* independent of being *on CIA team* (see Figure 10c). For judgments of gender from height, there were three possible responses: *more likely male*, *more likely female*, or *about the same*. A significant difference was again found in responses levels between the two scenarios ( $\chi^2(2)=7.3$ ,  $p<.05$ ). This time, responses of “about the same” (indicating base rate ignorance) were higher for the *selected by height* causal structure than the *selected by gender* causal structure, which is consistent with the prescription of the causal Bayesian norm that when selected by height, knowing the agent's *height* renders *gender* independent of being *on CIA team* (see Figure 10c).

We also analyzed the two factors separately, as well as their interaction, using a two-way ANOVA. We found no main effect of causal structure or and no main effect of judgment, but a highly significant interaction between them ( $\chi^2(2)=16.44$ ,  $p<0.0005$ ). Consistent with our predictions, base rate ignorance were significantly higher when the observed variable was situated between the variable to be judged and *on CIA team* in the causal model, whereas base rate use was significantly higher when the variable to be judged was situated between the observed variable and *on CIA team* in the causal model (see Figure 12). This indicated that nothing about the causal structure (*selected by gender* vs. *selected by height*) or the judgment (*gender from height* vs. *height from gender*) determines whether the base rate is used or ignored; rather, it is the interaction of the judgment with the scenario that determines whether the base rate is relevant. This demonstrates that people often use or ignore the base rate exactly as prescribed by our proposed causal Bayesian norm, and that ignoring the base-rate may often be a natural consequence of reasoning causally.

### Discussion

In several comparisons, we found strong effects of causal structure on judgments, demonstrating that people's use or neglect of the base rate varies according to what is rational under our proposed causal Bayesian norm. These results are not predicted by the heuristics and biases view, which proposed that people neglect the base rate when the individuating information seems more salient or more causally relevant than the base rate. In our questions, the base rate (70% under 5'7" or 70% female) remains constant across *scenarios*, yet we were able to induce use or neglect of the base rate by varying causal structure. The results also cast doubt on the hypothesis that people use a representativeness heuristic (Kahneman & Tversky, 1973) in social judgment tasks. A representativeness heuristic would judge  $P(\text{male} | \text{over } 5'7")$  by the degree to a person over 5'7" represents a typical male, ignoring the base rate of males on the team. Our results show that people generally use the low base rate of males in the *selected by gender* scenario appropriately, with the modal judgment being that a person over 5'7" on this team is more likely to be female than a person over 5'7" on other teams, contrary to the predictions of representativeness.

The interaction of causal structure with base rate use may be responsible for a host of base-rate neglect findings on social judgment types of tasks, such as in Kahneman and Tversky's (1973) lawyer-engineer problem, described in the introduction to this experiment. Assuming that personality is a causal influence on the career choice of lawyers and engineers, the appropriate causal model for this scenario, including the three phases of causal Bayesian inference, is depicted in Figure 9. We are told that the base rate of lawyers and engineers interviewed is  $P(\text{lawyer} | \text{interviewed}) = 70\%$ ;  $P(\text{engineer} | \text{interviewed}) = 30\%$  (in one variant), but this base rate cannot be used to update the CPT for the *career* variable, because it has a parent in the

model. The CPT for *career* only contains parameters for  $P(\textit{career} \mid \textit{personality})$ , not  $P(\textit{career})$ . One could accommodate this base rate by including a variable for *interviewed*. However, we are not told how the lawyers and engineers were selected for interview. It is therefore not clear how a variable representing *interviewed* should be connected causally to the rest of the model. If the *interviewed* variable is not causally connected, it is understandable that participants may ignore the base rate, assuming that whatever the causal connection is, learning the man's *personality*, a direct cause of *career*, renders the fact that the man was interviewed irrelevant to judging the man's *career*, just as learning the agent's *gender* rendered the fact that the agent was on the CIA team *selected by gender* irrelevant to judging the agent's *height*.

### General Discussion

Our experimental results suggest that people's judgments under uncertainty may be best understood in terms of causal Bayesian inference: approximately rational statistical inferences over mentally represented causal models. Results from four experiments support several distinct claims. First, people's judgments under uncertainty vary depending on the causal structure they believe to be underlying given statistics, and correspond well to the prescriptions of causal Bayesian inference when that framework makes clear predictions. Second, when provided with a clear causal model and statistics that can be mapped clearly onto that model, people typically use base rates appropriately, which includes ignoring base rates when a causal Bayesian analysis suggests they should be ignored. Finally, people make approximately rational judgments which cannot be explained using classical non-causal Bayesian norms (e.g., people rationally ignore the base rate of heights when judging height from gender in the CIA *selected by gender* scenario). In contrast to the heuristics and biases view, which casts people's ability to reason from

probabilistic information as highly suspect, all four experiments found that participants' modal judgments are rational by the standards of causal Bayesian inference.

While we have interpreted the results of our experiments as supporting the causal Bayesian hypothesis, other interpretations are possible. Our manipulations in Experiments 1-4 often make the newly developed versions of judgment problems seem easier, more salient or more engaging, which may lead to better performance. Even if our newly provided statistics are naturally more salient, our results are inconsistent with the notion that mere salience of statistics drives attention which drives usage. For instance, people actually used the false-positive statistic in Experiment 1 as often as they used the benign cyst statistic, they just tended to misinterpret it as  $P(\neg cancer | +M)$  instead of  $P(+M | \neg cancer)$ . We would argue that increased intelligibility of our new scenarios comes as a result of the change to a more natural causal structure, and we have attempted to control for confounds that could have led incidentally to better performance on these versions. Furthermore, in Experiment 4 both versions of the scenario were descriptively equivalent, hence arguably equally salient, therefore these experiments confirm that causal structure influences judgments independently of salience. Another account of our results could hold that judgment is guided by causal reasoning heuristics, which might often approximate causal Bayesian inference but in some cases fall short of that normative framework's full capacity. We do not see the use of heuristics as inconsistent with our view. Rather, we treat causal Bayesian reasoning as a rational method of inference that, like any computation, can be approximated with heuristics. Thus, although future studies may provide evidence for a more heuristic level of judgment, we expect that these heuristics will be better characterized as approximations to causal Bayesian inference than to classical statistical methods.

*Relation to the heuristics and biases view*

The heuristics and biases view previously identified causality as playing an important role in base-rate neglect (Ajzen, 1977; Tversky & Kahneman, 1980), but there are two ways in which our account is more compatible with the results presented here. First, researchers working in the heuristics and biases tradition predicted that “causal” (or otherwise salient) statistics would dominate non-causal (or non-salient) base rates. While this may seem close to our view, it runs directly counter to our main finding. We showed that supposedly “non-causal” and “non-salient” base rates (such as the rate of breast cancer and the proportion of cabs that are blue) are more likely to be used correctly when the other statistics given are more “causal” (i.e., when statistics that fit poorly into a causal model, such as a false alarm rate or a perceptual error rate, are replaced with statistics that clearly map onto parameters of the causal model, such as the base rate of an alternative cause).

Second, and more deeply, the heuristics & biases literature did not address the crucial rational function of causal knowledge in real-world judgment under uncertainty, but rather treated causal reasoning as a potentially distracting heuristic leading to judgment errors. In contrast, we view causal inference one of the foundations of people’s intuitive judgment ability, and we interpret effects of causality on judgment as core evidence for understanding how judgment works so well. We have attempted to explain how causal reasoning constitutes a rational system for making everyday judgments under uncertainty – in an adaptive sense, more rational than classical statistical norms – and we have shown how people’s judgments under uncertainty, and deviations from normative statistical answers, may reflect sophisticated causal reasoning abilities. Rather than trying to identify the factors that induce or reduce errors such as base-rate neglect, as in Bar-Hillel (1980), we have explained how a rational inference engine can

yield seemingly irrational judgments when people assume a different causal structure from the experimenters or are presented with statistical data that do not correspond to model parameters.

*Relation to the natural frequency hypothesis*

The natural frequency hypothesis (Gigerenzer & Hoffrage, 1995) claims that natural frequencies are the only statistical data that can be handled by the cognitive engine people have evolved for statistical inference. Probabilities and relative frequencies (i.e., percentages) are described as recently invented statistical formats that fail to activate people's natural statistical abilities. Because the hypothesis does not address how people reason with explicit probabilities, it cannot account for results that show good performance on problems involving probabilities or relative frequencies. Since people have been shown to be adept at reasoning with probabilities and percentages under the right circumstances, in the experiments presented here and in other studies (e.g., Bar-Hillel, 1980; Peterson & Beach, 1967) the natural frequency hypothesis does not seem to provide a general account of when and why judgments succeed or fail. Furthermore, because it relies on a purely statistical framework, the natural frequency hypothesis has significant limitations in its ability to account for real-world judgment under uncertainty, where there are often far too many variables and far too few prior observations to make valid judgments by natural frequencies alone.

In arguing against the natural frequency hypothesis, we do not mean to imply that natural frequencies are not useful. On the contrary, they are an extremely important source of input that can be used for updating parameters of causal models or for computing proportions when only two variables are of interest. We also do not object to the claim that people are skilled at reasoning with natural frequencies; the data clearly show they are. But rather than conclude that evolution has only equipped us with a natural frequency engine, we would argue that people do

better on these tasks because the natural frequency format makes the task simpler; it highlights nested sets that can be used in a calculation of proportions. Thus, performance improvements alone are not convincing evidence that people are naturally better at reasoning about frequencies than probabilities. Unlike most natural frequency experiments, our experiments were carefully controlled such that both conditions required equally complex calculations. Thus we can be more confident that the differences in performance between the conditions are due to the causal content of the scenarios rather than their mathematical form.

*Explaining other apparent errors of judgment under uncertainty*

We anticipate that a number of other “fallacies” in the judgment literature may be artifacts of attempting to analyze people’s causal judgments as approximations to traditional statistical methods, and may be more productively explained in terms of causal Bayesian reasoning. We do not claim this view can account for all cases where intuitive judgments appear to depart from classical statistical norms, or even for all cases of base-rate neglect, but there are several important classes of judgments where it appears to offer some useful insights.

One such class of judgments are “causal asymmetries”, where people more readily infer effects from causes, as in judgments of  $P(E|C)$ , than causes from effects, as in judgments of  $P(C|E)$ . For example, Tversky and Kahneman (1980) report that people expressed more confidence in judging a man’s weight from his height than a man’s height from his weight. In the context of causal Bayesian reasoning, this asymmetry is understandable:  $P(E|C)$  is a parameter of a causal model, which should be available directly from causal domain knowledge, whereas  $P(C|E)$  requires a Bayesian computation that depends on knowledge of  $P(E|C)$ , and thus should be more difficult to judge. Intuitively, when judging  $P(C|E)$ , one must often consider many potential causes of  $E$  and integrate over the possible states of these causes. No such

complexity is involved in judging  $P(E|C)$ , which can be obtained directly from the CPT of a causal model.

Another judgment phenomenon that our framework addresses is the difficulty people have with Simpson's paradox. An important version of this paradox is characterized by  $P(E|C, K) \geq P(E|\neg C, K)$  for all sub-populations ( $K$ ), but  $P(E|C) < P(E|\neg C)$  when the populations are combined into one; in each of the subpopulations,  $C$  seems to cause  $E$ , but overall,  $C$  seems to prevent  $E$ . Waldmann & Hagmayer (1996) showed that if people believe the co-factor ( $K$ ) to be a cause of  $E$ , they correctly condition on  $K$  when inferring the strength and direction of the contingency (in this case, people who believe  $K$  is a cause of  $E$  would conclude that  $C$  causes, rather than prevents,  $E$ , because for a given  $K$ ,  $C$  makes  $E$  more likely). However, the study did not address what factors account for the degree to which the situation seems paradoxical.

Our proposal suggests that Simpson's paradox seems paradoxical because people generally interpret statistical data as parameters of a causal model. If people believe a causal link exists between  $C$  and  $E$ , people will interpret the contingency statistic,  $P(E|C) < P(E|\neg C)$ , to be describing the parameter of that single causal link. For example, consider the following statistics: overall, those who use sunscreen more often are more likely to get skin cancer, but for both sunbathers and non-sunbathers, those who use sunscreen more often are less likely to get skin cancer. Since we know that a causal link exists between sunscreen and skin cancer, we naturally interpret the statistic as a parameter describing the causal power of that link. In this case, the sense of paradox is driven by the impression that the power of sunscreen to cause cancer is reversed (from preventive to generative) when the populations are combined. This paradox does not occur for the following statistics: overall, those who wear sunglasses more often are more

likely to get skin cancer, but for both sunbathers and non-sunbathers, those who wear sunglasses more often are no more likely to get skin cancer. Because we know that sunglasses cannot causally influence skin cancer, we do not interpret the contingency to be describing the causal link. There is no paradox because the power of sunglasses to cause cancer is not being changed when the population is combined; rather, it is clear that the apparent contingency results from a common cause (sun exposure). It is the prior knowledge of the preventive link between sunscreen and skin cancer that differentiates these two examples; when a causal or preventive link is known to exist between two variables, people naturally interpret the statistic to be describing the power of that causal link. Simpson's paradox becomes strongest when the causal link between the two variables is well known, but the common cause is not readily apparent. For example, people who eat vegetables regularly are more likely to get cancer, yet for every age group, people who eat vegetables regularly are less likely to get cancer. The paradox is dissolved by discovering that older people eat more vegetables, thus old age is a common cause of both cancer and eating vegetables regularly.

### *Learning Structure from Statistical Data*

Causal Bayesian inference represents a method of combining statistical data with prior knowledge to make judgments. In the related area of learning causal structure, an active debate currently exists between views that emphasize the importance of statistical data versus prior knowledge. Glymour and Cheng's (1998) approach to causal structure induction seeks to formalize methods that enable both causal structure and parameters to be learned from statistical data alone. Waldmann (1996) argues that statistics alone are not enough to explain learning, and demonstrates that prior knowledge of causal directionality and causal relevance can affect learning causal structure from data. Taking this idea further, Ahn and Kalish (2000) argue that

prior mechanism knowledge, including knowledge of intermediate causes in a chain, is crucial for learning causal structure, and especially for resolving ambiguous correlations. Tenenbaum and Griffiths (2001, 2003) propose a model that synthesizes the roles of prior causal knowledge and statistical data, in which knowledge serves to constrain the space of possible causal structures, and that data can then be used to favor one structure over another. Our approach to judgment also calls for a synthesis between prior knowledge and statistics, but like Ahn and Kalish (2000), our experiments suggest that understanding how a causal mechanism works may be crucial to interpreting statistics that describe it.

#### *Deterministic Mechanisms and Randomly Occurring Causes*

Another way to interpret our results is in terms of a bias towards deterministic mechanisms. In the original mammogram and cab problems, the statistics given implied that the mechanisms were fundamentally stochastic, randomly generating positive mammograms 9.6% of the time with no cause, or randomly causing the witness to make a mistake 20% of the time. A number of studies have cast doubt on people's abilities to comprehend randomness, including such well-known phenomena as the gambler's fallacy, the hot-hand fallacy (Gilovich, Vallone, & Tversky, 1985), and the law of small numbers (Tversky & Kahneman, 1971). In experiments 1-3, as part of clarifying the causal structure of the scenario, we moved the main source of randomness from the efficacy of the mechanism to the presence or absence of other causal variables. It could be that people are good at reasoning about nondeterministic scenarios when the main source of nondeterminism is in the random occurrence of causal variables, but they find it less natural to reason about mechanisms randomly failing (unless those failures can be attributed to some external factor, at which point the source of randomness becomes the presence of a cause that deterministically disables the mechanism). The notion that causal reasoning may be

accomplished by modeling deterministic mechanisms, with indeterminism introduced through uncertainty about the presence of hidden causal variables, has recently been proposed in both the artificial intelligence (Pearl, 2000) and psychological literatures (Schulz, Sommerville, & Gopnik, in press; Luhmann & Ahn, in press).

### *Making Statistics Easy*

People clearly have natural abilities to interpret statistics, yet they are often poor at interpreting published statistical data. Previous researchers have argued that we can leverage people's known natural abilities to teach them how to interpret published statistics. Sedlmeier and Gigerenzer (2001), for instance, present evidence that people can be taught to recast probabilities or relative frequencies, expressed as percentages, as natural frequencies, expressed as whole numbers, which improves correct response rates. However, if one does not hypothesize that people have a specialized cognitive engine for using natural frequencies, this effect could be seen as the result of mathematically simplifying a difficult problem. Just as one can break down a multi-digit multiplication problem into several single-digit multiplications added together, one can break down a probability question by considering nested subsets of a large number of individuals. The problem certainly becomes easier, but one need not hypothesize a special cognitive engine to explain it.

Our causal Bayesian hypothesis suggests a different approach to making statistics easy: replacing statistics that fit poorly into a causal model with statistics that correspond directly to causal model parameters. Furthermore, by embedding the statistics within a causal model, people may be able to understand the world better than they would with statistics alone, regardless of whether they are probabilities or natural frequencies. Consider the contrast in insight between the false-positive and benign cyst mammogram scenarios of Experiment 1 if one attempts to

generalize beyond the very restricted circumstances given in the problem setup. Imagine a woman who gets a positive mammogram, but hears that it can be unreliable so she decides to get a second mammogram. If that second mammogram also comes back positive, how much more confident should we be that she has cancer?

- The statistics in the false-positive problem suggest that she should be much more concerned after the second mammogram comes back positive: 6% of women without cancer will receive a positive mammogram, therefore 0.36% will test positive twice, assuming the second mammogram result is independent of the first. The chance of having cancer given

two positive mammograms, then, is  $\frac{2\%}{2\% + 98\% \times 0.36\%} = 85\%$ .

- In contrast, the causal structure described by the benign cyst mammogram scenario of Experiment 1 suggests that the unreliability of the mammogram apparent in the false-positive scenario is an illusion. The mammogram reliably detects cancer, but it also reliably detects benign cysts, and if a woman has a benign cyst she will get a positive mammogram the second time as well. In this scenario, two positive mammograms is no more diagnostic of

cancer than one positive mammogram. The answer remains  $\frac{2\%}{2\% + 98\% \times 6\%} = 25\%$ .

It may seem strange that supposedly equivalent statistics lead to greatly different inferences. This occurs because both sets of statistics are mere reflections of a much more complex underlying generative process. For instance, you may believe that any benign cyst has a 50% chance of being detected, or you may believe that only 50% of benign cysts are dense enough to be detected, but those that are dense enough will be detected every time. This second situation can be modeled causally by adding variables to represent the size and density of the cyst or tumor, and then specifying a threshold at which it is large enough to be detected deterministically. Of

the views we have considered, we believe that only the causal Bayesian hypothesis can account for how people extrapolate meaning by going beyond the statistics to represent a causal model, and using the model to make new inferences that are under-determined with statistics alone.

The most intuitively valuable statistics are those that correspond transparently to parameters of known causal relationships. However, for some situations, the true causal structure may not be obvious to people. In these cases, one should explain to people the true causal structure, as well as provide statistics that map onto that structure. For example, consider the statistic that patients are more likely to survive after being treated by doctor B than by doctor A (from Bar-Hillel, 1990). One could easily get the impression that doctor A provides inferior care, unless one is specifically informed that doctor A specializes in life-threatening diseases, and doctor B does not. This new information invokes a causal structure in which a person's disease state causally influences both the choice of doctor and the likelihood of survival. With this new structure, it is easy to see that the low survival rate of doctor A's patients may be due solely to a higher base rate of patients with life-threatening diseases, and hence the quality of care may be the same or even better than that of doctor B. But with the wrong causal structure in mind, people could easily and understandably jump to false and dangerous conclusions.

### Conclusion

The need to make intuitive statistical judgments is a pervasive fact of life in human society. But if we relied only on purely statistical information, we would be in dire straits, as the remarkable flexibility, success, and inductive potential of common sense would be impossible. Fortunately, our physical, biological and social environments are causally structured, and our intuitive theories of the world are often – but not always – sufficient to capture the most relevant structures for enabling appropriate causal Bayesian inferences. In experimental studies, if we

present participants with a clear causal structure and statistics that clearly map onto that structure, we can nearly eliminate traditional judgment errors such as base-rate neglect and dramatically boost the incidence of correct Bayesian reasoning. Those who have a stake in improving statistical reasoning in complex, everyday settings – scientists, educators, doctors, advertisers, politicians, and many others – could do well to follow the same approach in communicating their questions, their data, and their conclusions to the lay public.

## References

- Ahn, W., & Kalish, C. (2000). The role of covariation vs. mechanism information in causal attribution. In R. Wilson, & F. Keil (Eds.) *Cognition and explanation*. Cambridge, MA: MIT Press.
- Ahn, W. (1999). Effect of Causal Structure on Category Construction. *Memory & Cognition*, 27, 1008-1023.
- Anderson, J. (1990). *The Adaptive Character of Thought*. Hillsdale, NJ : Erlbaum Associates.
- Ajzen, I. (1977). Intuitive Theories of Events and the Effects of Base-Rate Information on Prediction. *Journal of Personality and Social Psychology*, 35 (5), 303-314.
- Bar-Hillel, M. (1990). Back to Base Rates. In RM Hogarth (Ed), *Insights in Decision Making: A Tribute to Hillel J. Einhorn*, 200–216. Chicago: University of Chicago Press.
- Bar-Hillel, M. (1980) The base-rate fallacy in probability judgments. *Acta Psychologica*, 44, 211-233.
- Birnbaum, M.H. (1983). Base Rates in Bayesian Inference: Signal Detection Analysis of the Cab Problem. *American Journal of Psychology*, 96, 85-94.
- Cheng, PW (1997). From covariation to causation: A causal power theory. *Psychological Review*, 104(2), 367-405.
- Cosmides, L. and Tooby, J. (1996). Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty. *Cognition*, 58, 1-73.
- Eddy, D. M. (1982). Probabilistic reasoning in clinical medicine: Problems and opportunities. In D. Kahneman, P. Slovic, & A. Tversky (Eds), *Judgment under uncertainty: Heuristics and biases*, 249-267. Cambridge: Cambridge University Press.

- Friedman, N., Linial, M., Nachman, I., and Pe'er D. (2000). Using Bayesian Networks to Analyze Expression Data. *Proceedings of the Fourth Annual International Conference on Computational Molecular Biology (RECOMB 2000)*, 127-135.
- Gigerenzer, G. & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological review*, 102 (4), 684-704.
- Gilovich, T., Vallone R., & Tversky, A. (1985). The Hot Hand in Basketball: On the Misperception of Random Sequences. *Cognitive Psychology*, 17, 295-314.
- Glymour, C. (2003). Learning, prediction and causal Bayes nets. *Trends in Cognitive Sciences*, 7, 43-48.
- Glymour, C. (2001). *The Mind's Arrows: Bayes Nets and Graphical Causal Models in Psychology*. Cambridge, MA: The MIT Press.
- Glymour, C., & Cheng, P. (1998). Causal Mechanism and Probability: A Normative Approach. in M. Oaksford and N. Chater (Eds.), *Rational Models of Cognition*. Oxford: Oxford University Press.
- Gopnik, A., Glymour, C., Sobel D., Schulz L., Kushnir, T., & Danks, D. (2004). A theory of causal learning in children: Causal maps and Bayes-Nets. *Psychological Review*, 111 (1), 3-32.
- Gopnik, A., & Glymour, C. (2002). Causal maps and Bayes nets: A cognitive and computational account of theory-formation, in S. Stich, P. Carruthers (Eds.), *The Cognitive Basis of Science*, 117-132. New York: Cambridge University Press.
- Gopnik, A., & Sobel, D. (2000). Detectingblickets: How young children use information about novel causal powers in categorization and induction. *Child development*, 71 (5), 1205-1222.

- Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology* 51(4), 285-386.
- Jordan, M. I. (Ed.) (1999). *Learning in Graphical Models*. Cambridge, MA: MIT Press.
- Kahneman, D. & Tversky, A. (1973). On the psychology of prediction. *Psychological Review* 80 (4), 237-251.
- Kahneman, D. & Tversky, A. (1972). On prediction and judgment. *Oregon Research Institute Bulletin* 12 (4).
- Krynski, T.R., Tenenbaum, J. B.. (2003). The Role of Causal Models in Reasoning Under Uncertainty. *Proceedings of the 25th Annual Conference of the Cognitive Science Society*.
- Luhmann, C. C., & Ahn, W. (in press). The meaning and computation of causal power: A critique of Cheng (1997) and Novick and Cheng (2004). *Psychological Review*.
- Lyon, D. & Slovic, P. (1976). Dominance of accuracy information and neglect of base rates in probability estimation. *Acta Psychologica*, 40, 287-298.
- McKenzie, C. R. M. (2003). Rational models as theories -- not standards -- of behavior. *Trends in Cognitive Sciences*, 7, 403-406.
- Nisbett, R. E. & Borgida, E. (1975). Attribution and the psychology of prediction. *Journal of Personality and Social Psychology*, 32, 932-943.
- Oaksford, M. & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review*, 101, 608-631.
- Oatley, G. & Ewart, B. (2003). Crimes analysis software: 'pins in maps', clustering and Bayes net prediction. *Expert Systems with Applications*, 25 (4), 569-588.
- Pearl, J. (2000). *Causality: models, reasoning, and inference*. New York: Cambridge University Press.

- Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Francisco: Morgan Kaufmann.
- Peterson, C. & Beach, L. R. (1967). Man as an intuitive statistician. *Psychological Bulletin*, 68 (1), 29-46.
- Rehder, B. (2003). Categorization as causal reasoning. *Cognitive Science*, 27, 709-748.
- Russell, S. J. & Norvig, P. (2003). *Artificial Intelligence: A Modern Approach, 2nd Edition*. Englewood Cliffs, NJ: Prentice-Hall.
- Schulz, L. E., Sommerville, J., & Gopnik, A. (in submission). God does not play dice: Causal determinism and children's inferences about unobserved causes. *Child Development*.
- Sedlmeier, P. & Gigerenzer, G. (2001). Teaching Bayesian reasoning in less than two hours. *Journal of Experimental Psychology: General*, 130, 380-400.
- Shepard, R. N. (1987). Towards a universal law of generalization for psychological science. *Science*, 237, 1317-1323.
- Slooman, S.A., & Lagnado, D. (2005). Do we “do”? *Cognitive Science*, 29, 5-39.
- Slovic, P., Fischhoff, B., and Lichtenstein, S. (1976). Cognitive processes and societal risk taking. In J. S. Carol and J. W. Payne (Eds.), *Cognition and Social Behavior*. Hillsdale, NJ: Erlbaum.
- Spiegelhalter, D., Dawid, P., Lauritzen, S., Cowell, R. (1993). Bayesian analysis in expert systems. *Statistical Science*, 8, 219-282.
- Spirites, P., Glymour, C., and Scheines, R. (1993). *Causation, Prediction, and Search*. New York: Springer-Verlag.
- Steyvers, M., Tenenbaum, J., Wagenmakers, E.J., Blum, B. (2003). Inferring causal networks from observations and interventions. *Cognitive Science*, 27, 453-489.

- Tenenbaum, J. B. & Griffiths, T. L. (2001) Structure learning in human causal induction. *Advances in Neural Information Processing Systems*. Leen, T., Dietterich, T., and Tresp, V., Cambridge, MIT Press, 2001, 59-65.
- Tenenbaum, J. B. & Griffiths, T. L. (2003). Theory-based causal inference. *Advances in Neural Information Processing Systems 15*. Becker, S., Thrun, S., & Obermayer. (Eds). Cambridge, MIT Press, 35-42.
- Tversky, A. & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin*, 1971, Vol. 76, No. 2. 105-110.
- Tversky, A. & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185, 1124-1131.
- Tversky, A. & Kahneman, D. (1980). Causal schemas in judgments under uncertainty. *Progress in social psychology*, 49-72, M. Fishbein (Ed.). Erlbaum.
- Villejoubert, G. & Mandel, D. R. (2002). The inverse fallacy: An account of deviations from Bayes's theorem and the additivity principle. *Memory and Cognition* 30 (2), 171-178.
- Waldmann, M. R. (1996). Knowledge-based causal induction. In D. R. Shanks, K. J. Holyoak, & D. L. Medin (Eds.), *The psychology of learning and motivation, Vol. 34: Causal learning*, 47-88. San Diego: Academic Press.
- Waldmann, M. R. (2001). Predictive versus diagnostic causal learning: Evidence from an overshadowing paradigm. *Psychonomic Bulletin & Review*, 8, 600-608.
- Wasserman, L. (2004). *All of Statistics : A Concise Course in Statistical Inference*. Springer.

## Footnotes

<sup>1</sup> We should note that a popular explanation of the original mammogram problem suggests that people are confused by the given conditional probability  $P(+M | cancer)$ , and think it means  $P(cancer | +M)$ . This has been called the inverse fallacy (Villejoubert, & Mandel, 2002). We find this limited as an explanation, because people only seem to exhibit the inverse fallacy when they expect both probabilities to have roughly the same value. For instance, while people might agree that the probability of death in a plane crash is nearly 100% ( $P(death | plane\ crash) \approx 100\%$ ), they surely would not agree to the inverse: that death is almost always a result of plane crashes ( $P(plane\ crash | death) \approx 100\%$ ). Thus, it may be that people only confuse a conditional probability with its inverse when they expect the inverse to have the same value. It is this expectation, then, that needs to be explained.

<sup>2</sup> Another way to incorporate the false-positive statistic would be to adopt a different parameterization, other than the noisy-or model that describes the effects of one or more independent generative causes. The noisy-or parameterization is often appropriate when variables represent causal factors that can be present or absent, such as a variable representing whether or not a person has breast cancer; this is how we expect most people interpret the mammogram problem. But variables may also represent two or more distinct causally active states of a single factor. For instance, a variable could represent which of two medical conditions a person has; then, not having cancer would be equivalent to having some other condition with a potential power to cause positive mammograms (or other effects). In that case  $P(+M | \neg cancer)$  would fit clearly into the causal model, just as conditional probabilities of the same  $P(A | \neg B)$  form were represented in the causal models discussed earlier for the cancer-childless-stress levels scenarios shown in Figure 2. Yet for the standard mammogram problem, this would require extra

steps of hypothetical reasoning that we think might not occur to many participants in a rapid judgment task.

<sup>3</sup>The results are also significant by a  $\chi^2$  test, but because one of the cells of the expected distribution was less than 5, the  $\chi^2$  test is considered unreliable and therefore Fisher's Exact Test should be used.

<sup>4</sup>To be fair, classical Bayesian norms dictating that posterior judgments should change as priors change are not applicable to these scenarios, because the likelihoods are dependent on the priors. Although the dependence means the classical Bayesian norm is not applicable, the norm provides no means to determine whether this dependence exists, nor how to reason appropriately about it. In fact, we suspect many of the classic scenarios purportedly demonstrating judgment failures contain priors that are not independent of the likelihoods, yet researchers continue to hold people's judgments to inapplicable Bayesian norms requiring that changes in priors (base rates) should lead to changes in posterior judgments. Because the causal Bayesian framework provides not just a method of determining whether the priors and likelihoods are independent, but also a method of making judgments when they are dependent, it offers a more useful standard for rational judgment.

## Figure Captions

*Figure 1:* In our ideal model for causal Bayesian inference, judgment under uncertainty divides into three phases. (a) Given a task description, causal domain knowledge is used to construct a causal model (b) Available statistical data are used to set parameters values. (c) Judgments are made via Bayesian inference over the parameterized causal model. The particular task illustrated here corresponds to one of the conditions tested in Experiment 1, the “benign cyst” scenario.

*Figure 2:* Three causal structures that could be responsible for the same statistics relating *having cancer (C)*, *being childless (L)*, and *having high stress levels (S)*. The correct judgment for  $P(C|L,S)$  depends on which causal structure was actually responsible for generating the observed statistics.

*Figure 3:* The three phases of causal Bayesian inference for a more traditional version of the mammogram question, the “false positive” scenario tested in the Experiment 1. Under this interpretation of the causal structure of the false positive scenario, having breast cancer is the only explicitly represented cause of positive mammograms. In contrast, for the scenario depicted in Figure 1, there are two explicitly represented possible causes of positive mammograms: cancer or benign cysts.

*Figure 4:* Histogram of responses to Experiment 1. The correct answer was 25%. Responses were classified as *correct* (20%-25%), *base-rate neglect* ( $\geq 75\%$ ), *odds form* (33%), *base rate overuse* (2%), and *other*. For analysis purposes, responses of *odds form* and *base rate overuse* were grouped with *other*. A significant difference was found between *false positive* and *benign cyst* scenarios ( $\chi^2(2) = 6.28, p < .05$ ). Error bars represent the standard error of the normal approximation to the binomial distribution.

*Figure 5:* Histogram of responses to Experiment 2. The correct answer was 5.1%. Responses were classified as *correct* (5.1%), *base-rate neglect* ( $\geq 65\%$ ), and *other*. A significant difference was found between *false positive* and *benign cyst* scenarios (Fisher's Exact Test,  $p < .05$ ). Error bars represent the standard error of the normal approximation to the binomial distribution.

*Figure 6:* The three phases of causal Bayesian inference for the perceptual error scenario of Experiment 3. The causal model depicts one possible causal interpretation, in which the witness' judgment obeys signal detection theory, and the witness' error rate could be used to infer the discriminability of the colors.

*Figure 7:* The three phases of causal Bayesian inference for the faded paint scenario of Experiment 3.

*Figure 8:* Histogram of responses to Experiment 3. Error bars represent the standard error of the normal approximation to the binomial distribution. The difference between conditions was large and highly significant ( $\chi^2(2) = 11.55$ ,  $p < .005$ ).

*Figure 9:* Causal model for the lawyer-engineer problem. It is not clear how the *interviewed* variable should be connected causally to the rest of the model.

*Figure 10:* The three phases of causal Bayesian inference for the selected by gender scenario of Experiment 4.

*Figure 11:* The three phases of causal Bayesian inference for the selected by height scenario of Experiment 4.

*Figure 12:* Histogram of Experiment 4 results showing levels of base rate use compared to base rate neglect across *gender* vs. *height* scenarios. A two-way analysis of variance shows a significant interaction between the type of team (*selected by gender* vs. *selected by height*) and

the judgment required (*inferring gender from height vs. inferring height from gender*)

(  $\chi^2(2)=9.7$ ,  $p<.001$  ).

Figure 1

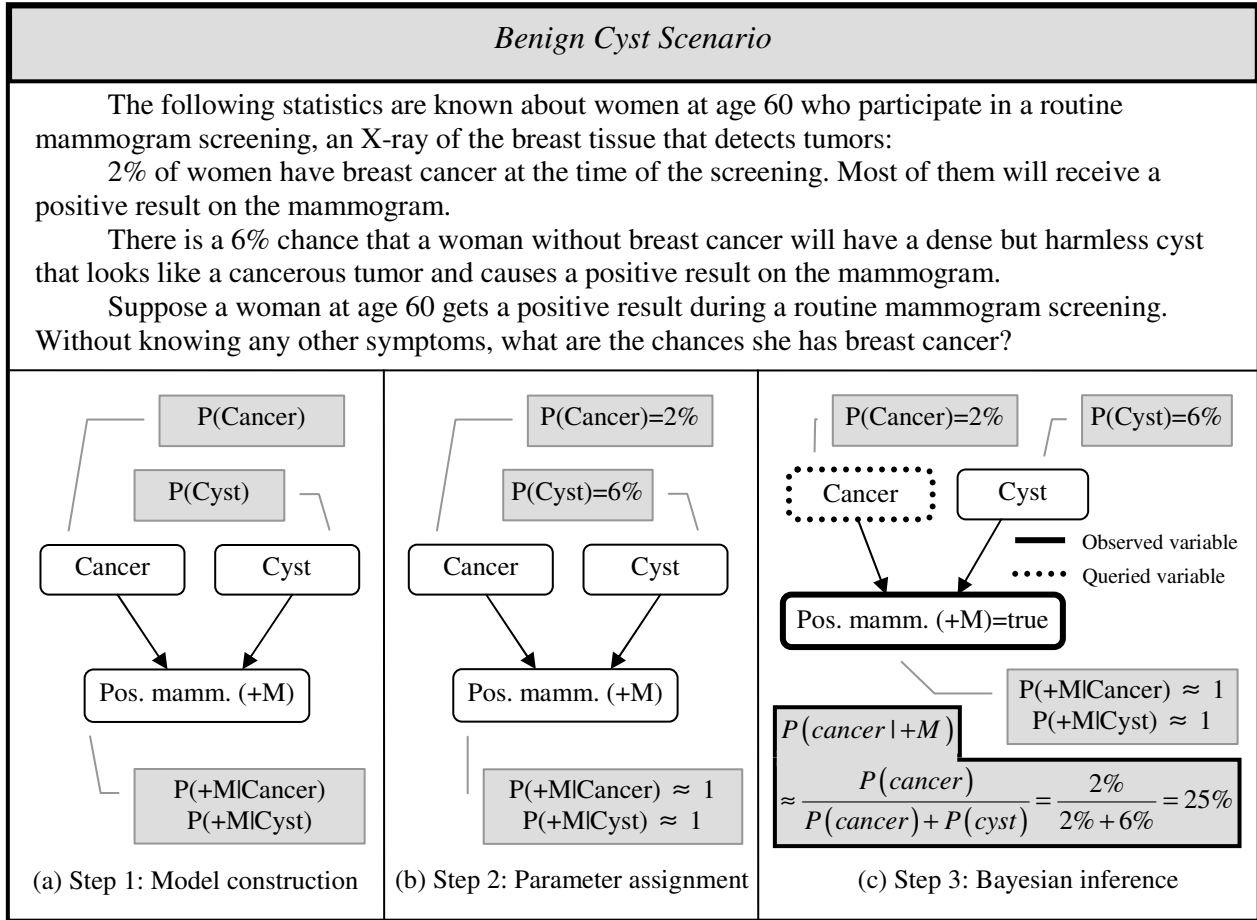


Figure 2

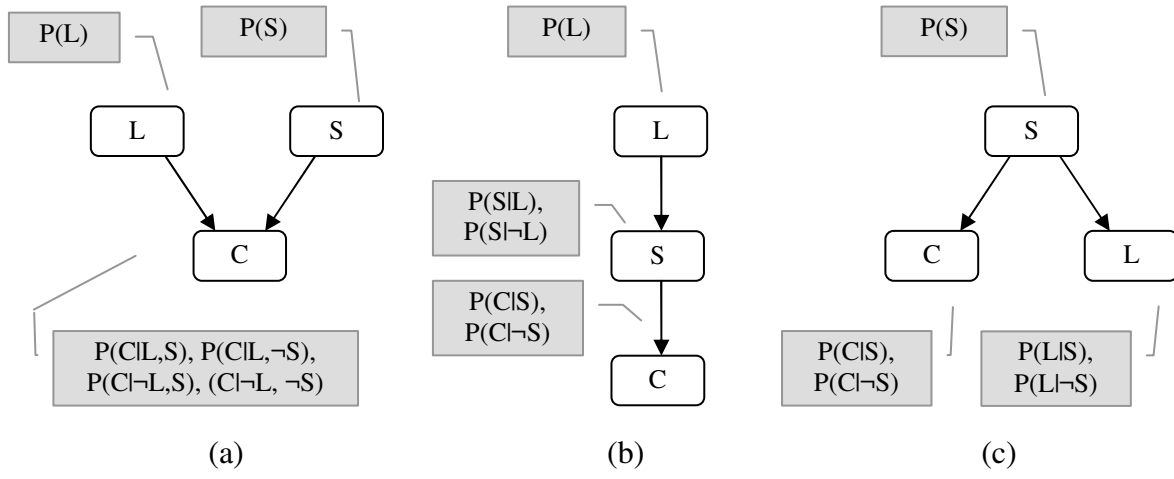


Figure 3

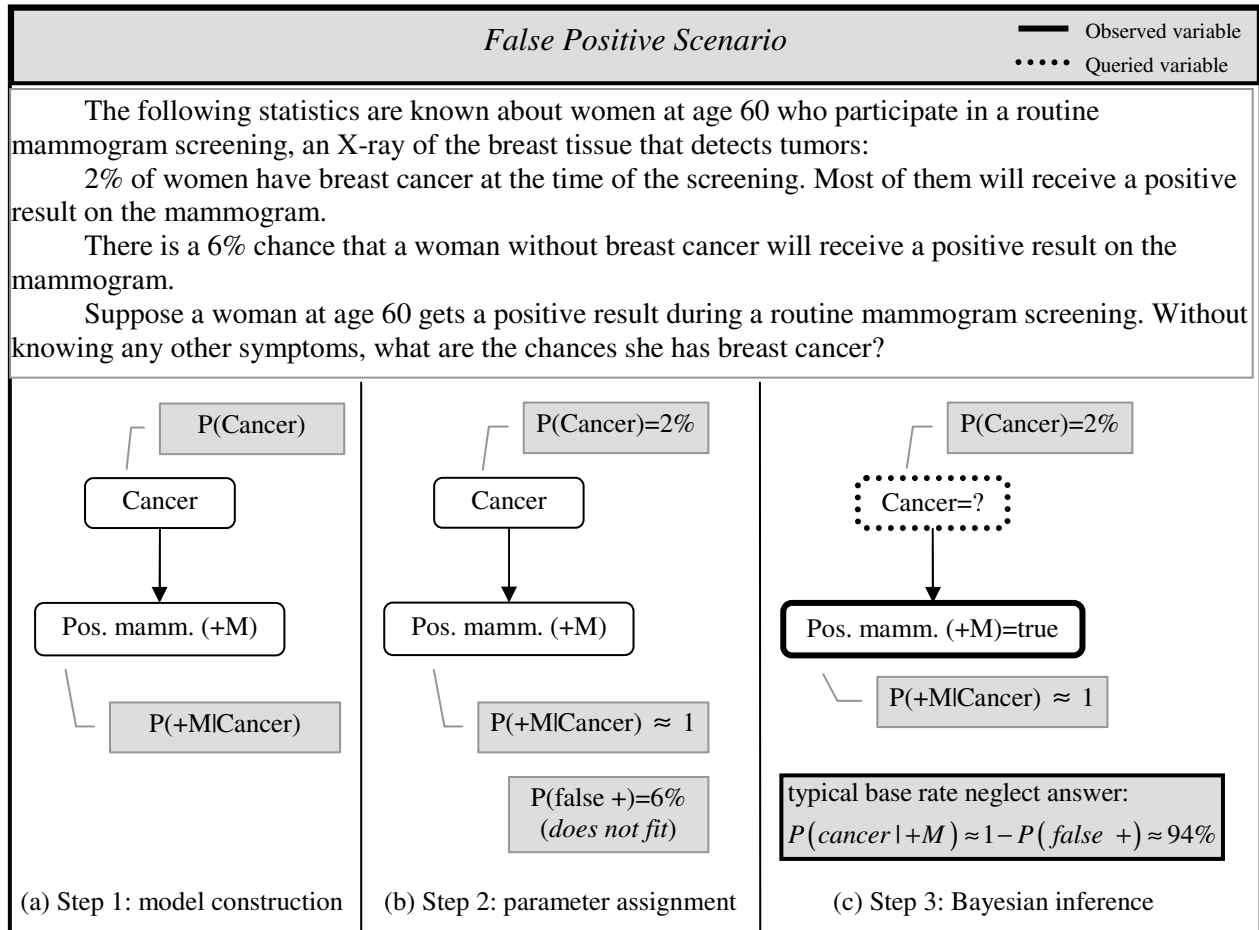


Figure 4

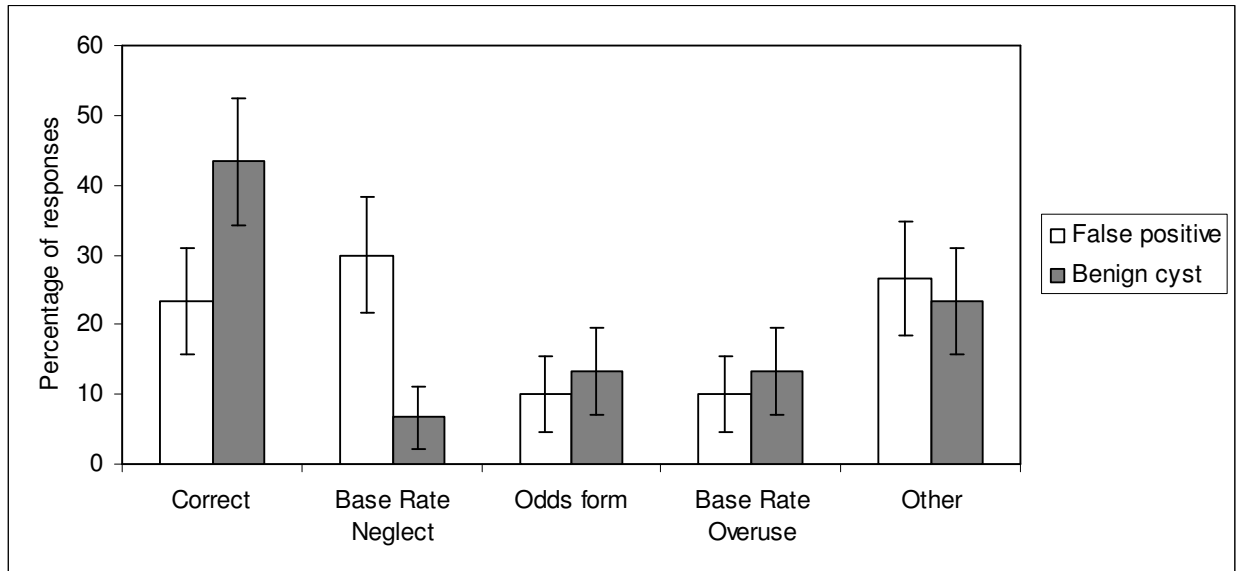


Figure 5

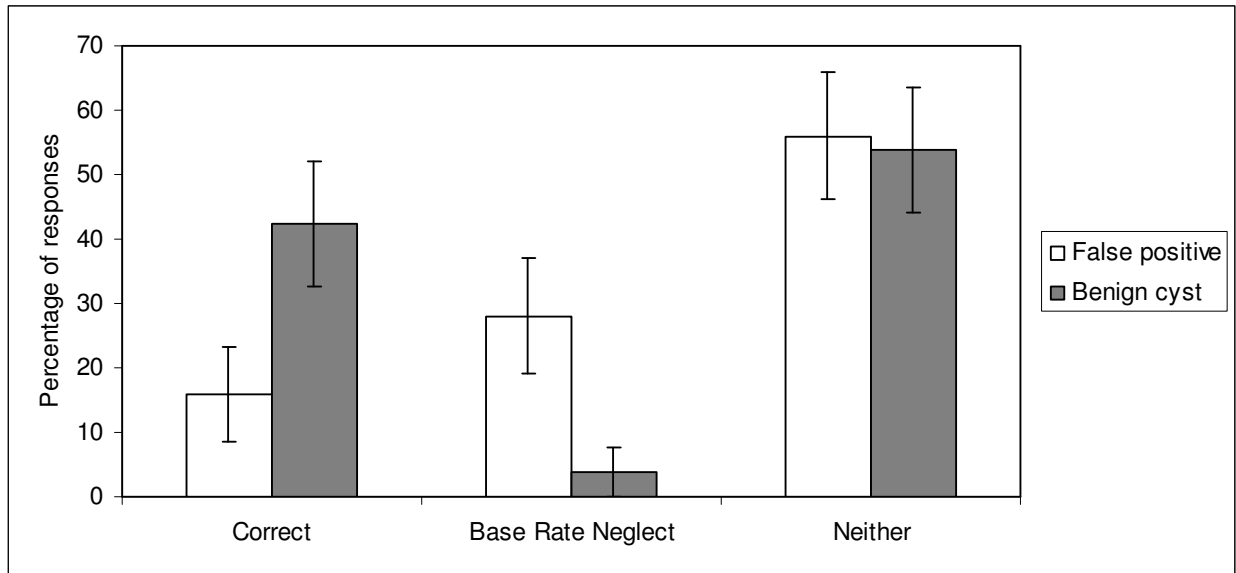


Figure 6

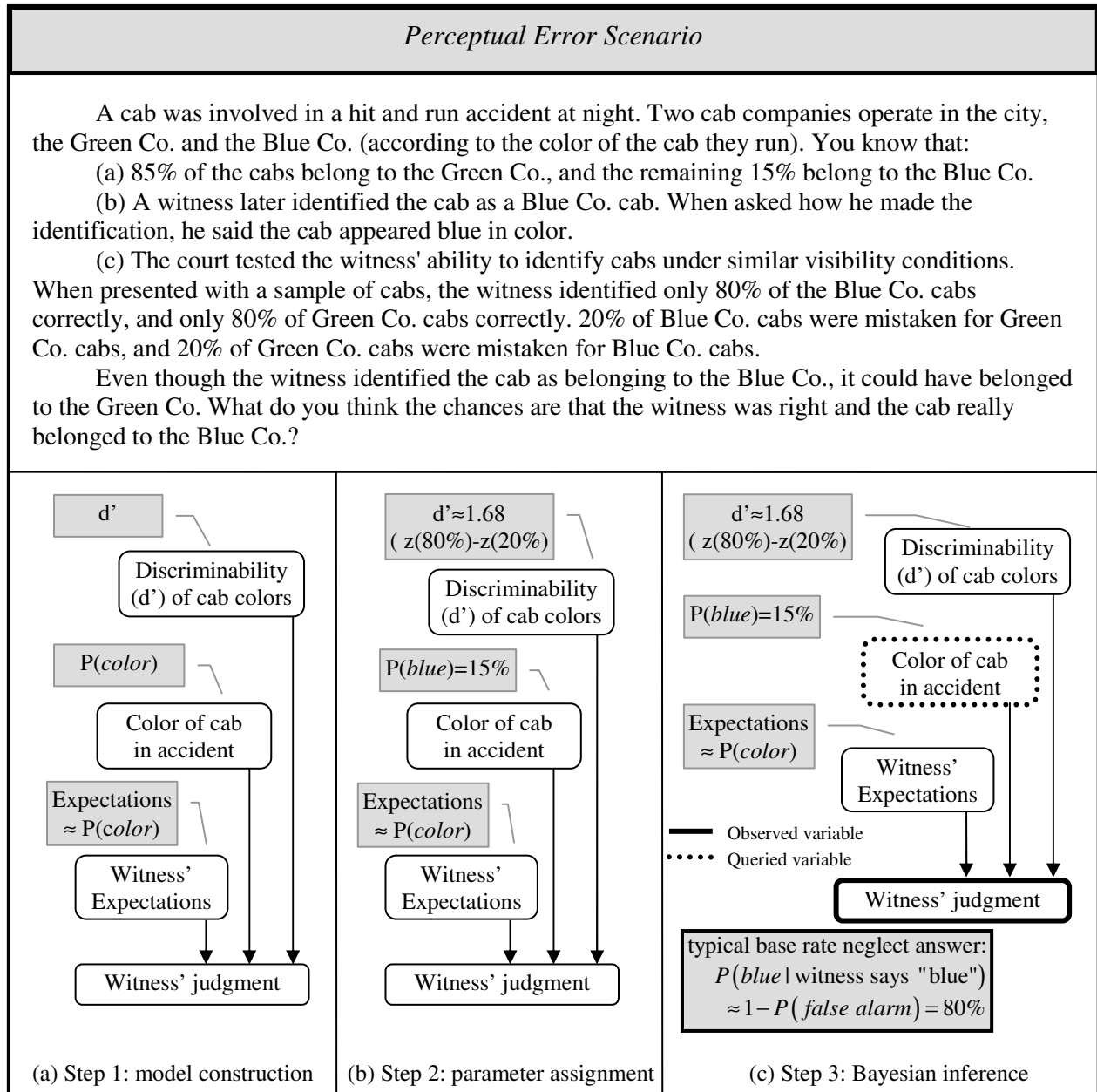


Figure 7

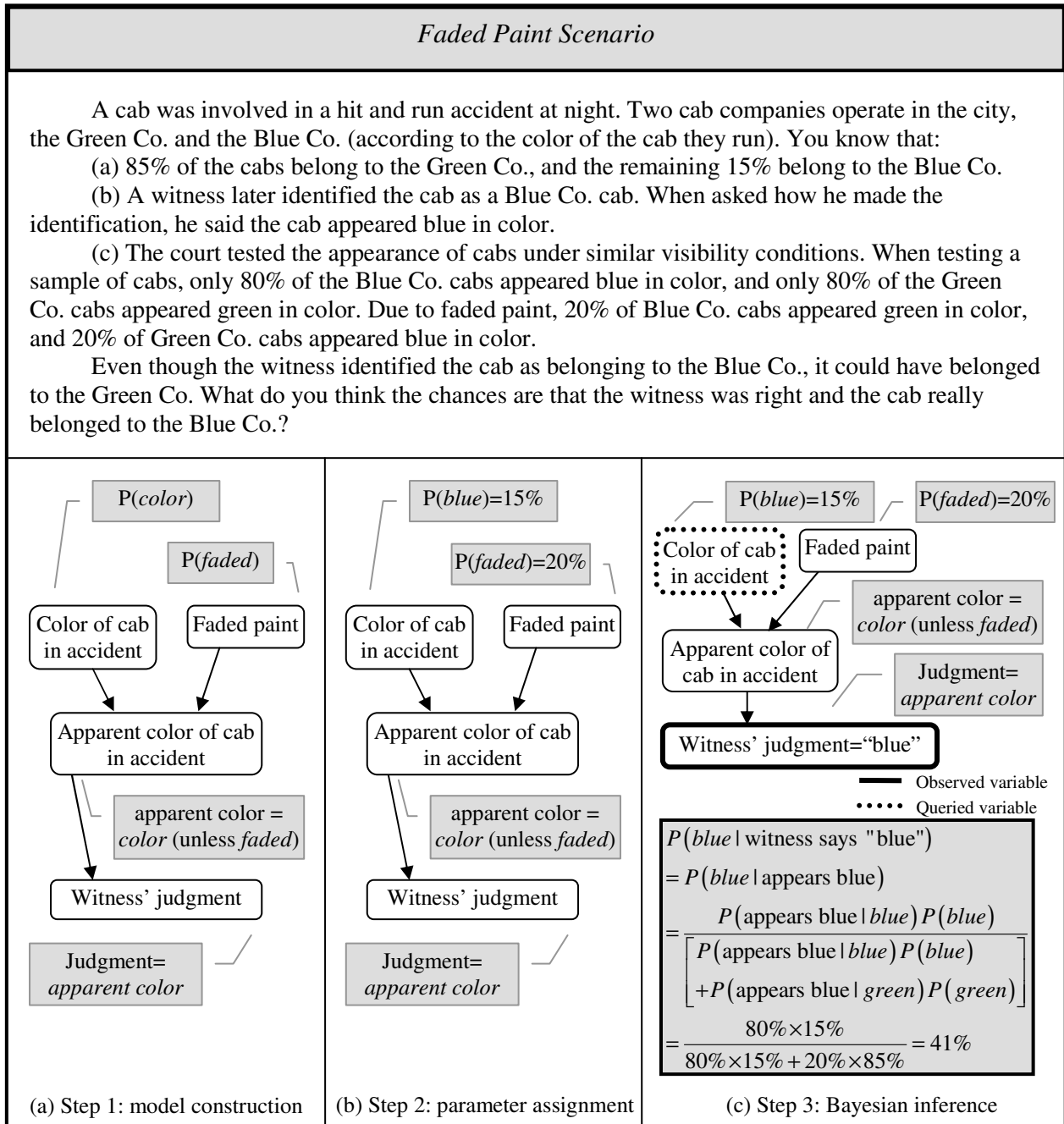


Figure 8

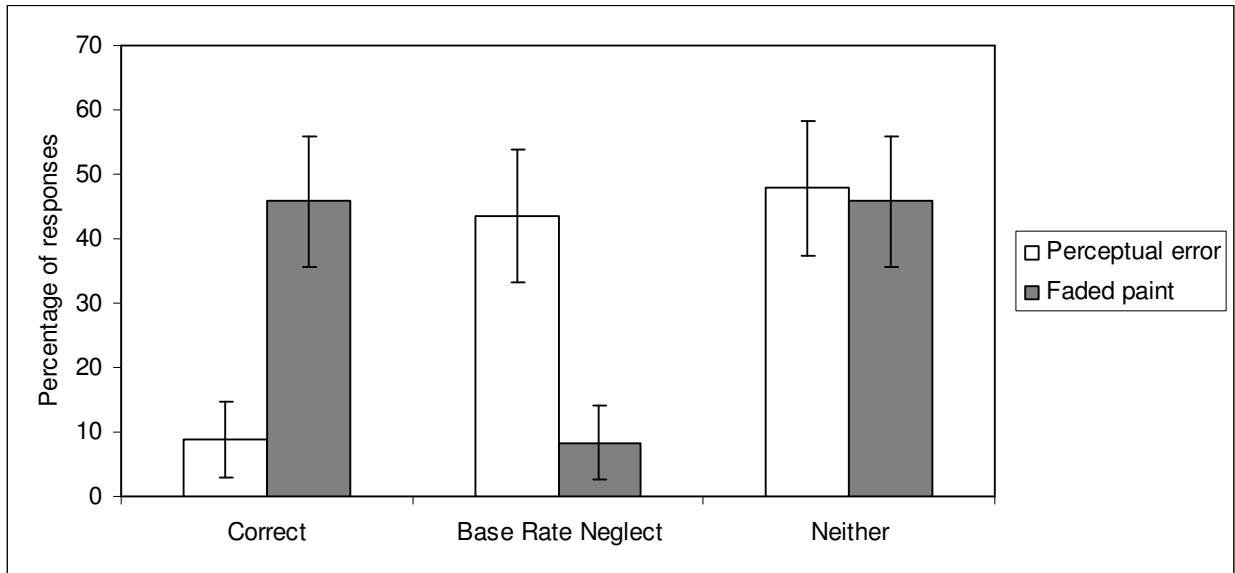


Figure 9

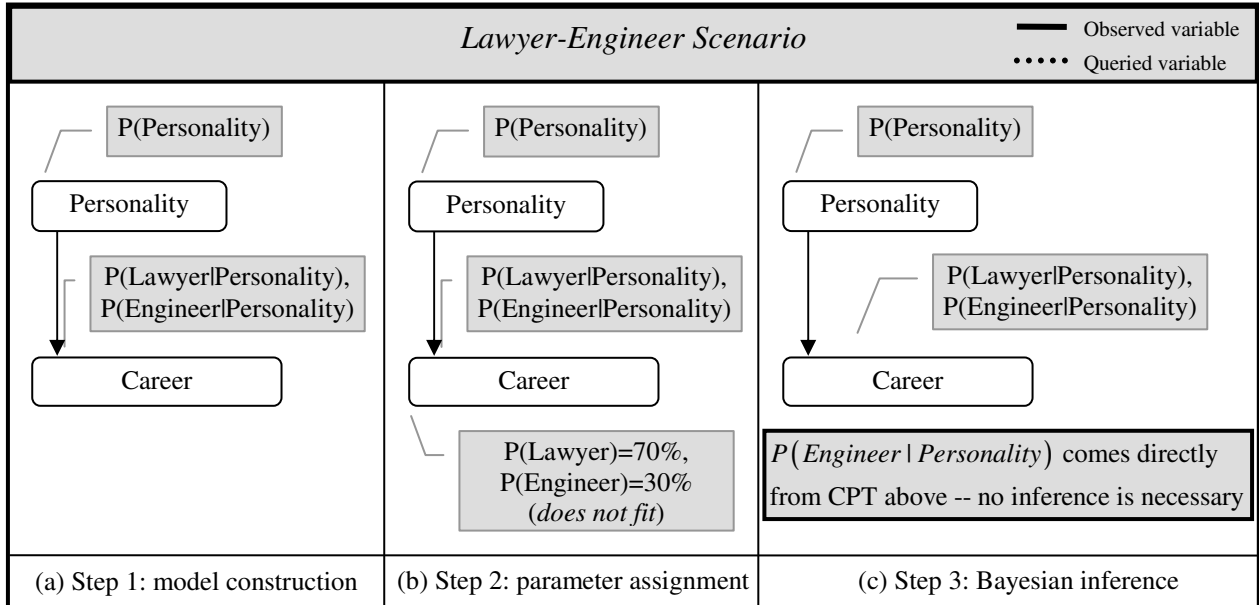


Figure 10

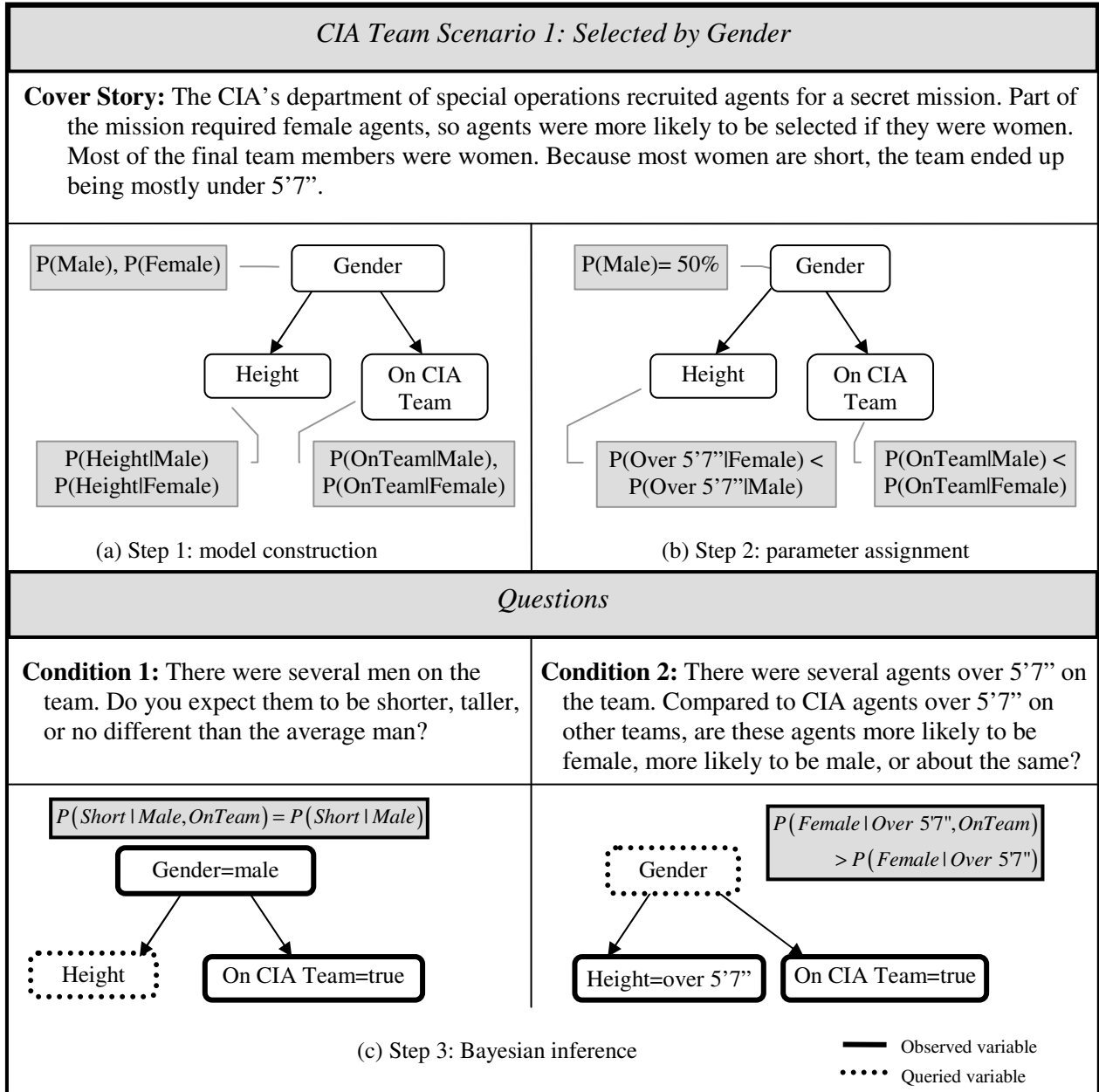


Figure 11

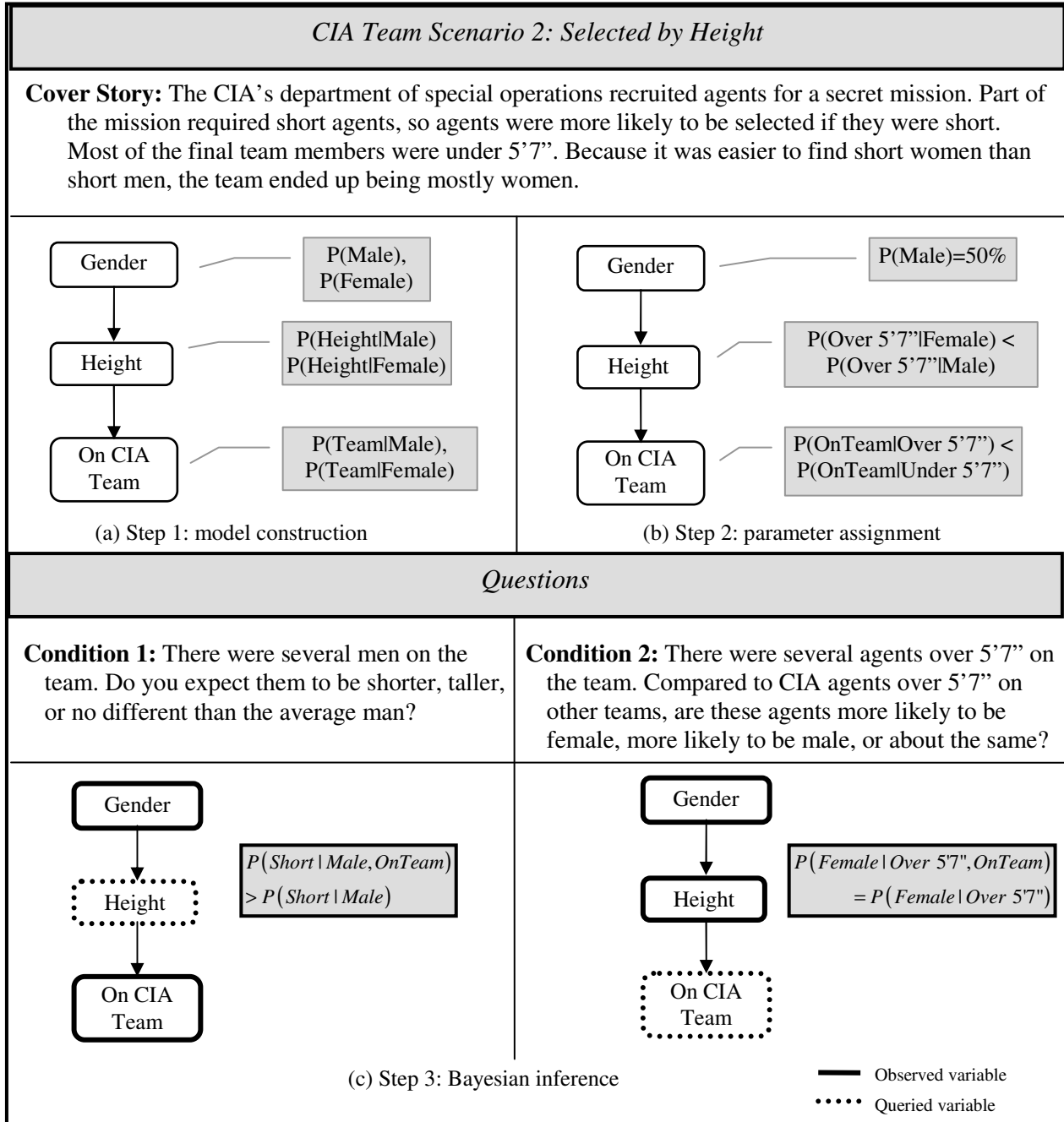
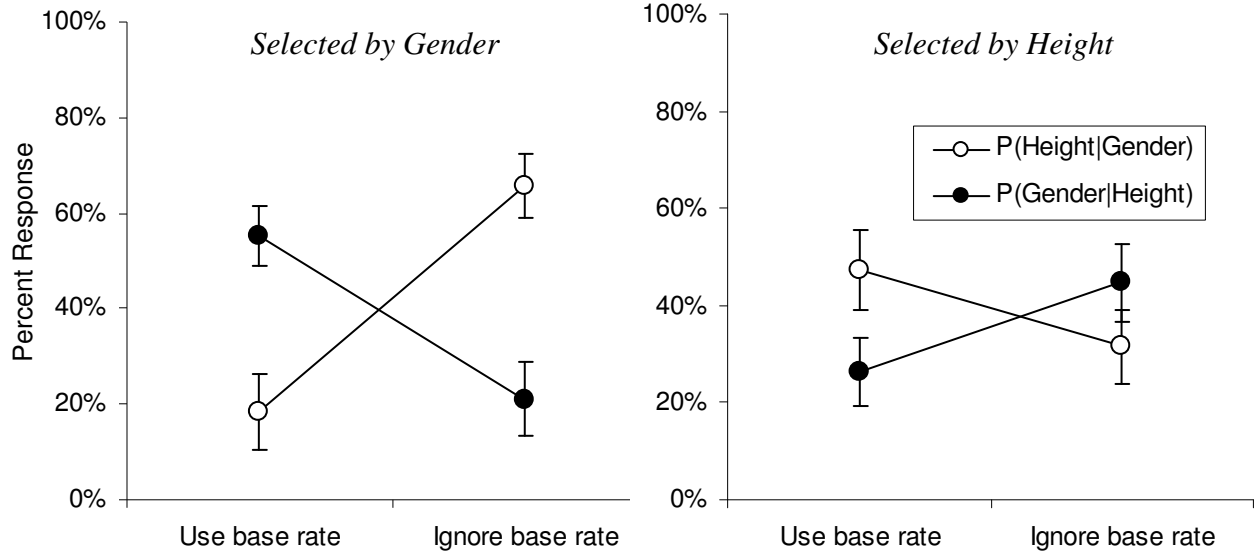


Figure 12



---