# Belief Polarization Is Not Always Irrational

Alan Jern, Kai-min K. Chang, and Charles Kemp
Carnegie Mellon University

Belief polarization occurs when 2 people with opposing prior beliefs both strengthen their beliefs after observing the same data. Many authors have cited belief polarization as evidence of irrational behavior. We show, however, that some instances of polarization are consistent with a normative account of belief revision. Our analysis uses Bayesian networks to characterize different kinds of relationships between hypotheses and data, and distinguishes between cases in which normative reasoners with opposing beliefs should both strengthen their beliefs, cases in which both should weaken their beliefs, and cases in which one should strengthen and the other should weaken his or her belief. We apply our analysis to several previous studies of belief polarization and present a new experiment that suggests that people tend to update their beliefs in the directions predicted by our normative account.

*Keywords:* belief polarization, belief revision, reasoning, rationality, Bayesian networks

*Supplemental materials:* http://dx.doi.org/10.1037/a0035941.supp

How much should we rely on prior beliefs when evaluating new evidence? Prior beliefs allow us to make sense of ambiguous evidence, but if weighted too heavily, they can cause us to ignore unexpected outcomes. In this article, we address the question of whether people rationally combine their prior beliefs with evidence. Researchers from several disciplines have explored this issue (Kahneman, Slovic, & Tversky, 1982; Kelly, 2008; Rabin & Schrag, 1999) and have documented cases in which human reasoning appears to be consistent with normative principles of belief revision (Gigerenzer, 1991; Koehler, 1993; Lopes & Ekberg, 1980; Rehder & Hastie, 1996) and cases in which it does not (Kahneman et al. 1982; Nisbett & Ross, 1980; Nisbett, Zukier, & Lemley, 1981; Pitz, Downing, & Reinhold, 1967). We focus on a phenomenon called belief polarization that is often described as an example of irrational behavior. We develop a normative account of belief revision, apply it to several classic studies of belief polarization, and present an experiment that suggests that some instances of belief polarization are consistent with our normative account.

Belief polarization occurs when two people with different prior beliefs observe the same data and subsequently strengthen their beliefs (Batson, 1975; Lord, Ross, & Lepper, 1979; Munro & Ditto, 1997; Plous, 1991). Figure 1 shows a simple example in which two doctors make a judgment about the same patient. The patient has either Disease 1 or Disease 2 and the two doctors initially disagree about the probability of each disease. The doctors observe the same piece of evidence—a cholesterol test result—and subsequently update their beliefs in opposite directions, both becoming more certain about their initial diagnoses. Many authors (Baron, 2008; Munro & Ditto, 1997; Ross & Anderson, 1982) have described belief polarization as a strictly irrational behavior. For instance, when discussing a classic study of belief polarization (Lord et al. 1979), Ross and Anderson (1982, p. 145) wrote that polarization is "in contrast to any normative strategy imaginable for incorporating new evidence relevant to one's beliefs."

Judging whether a behavior is irrational requires comparing it to a normative standard. We evaluate whether belief polarization is rational by using a formal analysis based on probabilistic inference, a normative standard for reasoning under uncertainty. We show that there are some situations in which polarization is indeed inconsistent with probabilistic inference, but others in which polarization emerges as a consequence of probabilistic inference. Multiple authors have presented analyses of belief polarization that are complementary to our own (Bullock, 2009; Gerber & Green, 1999; Jaynes, 2003). Some have shown that polarization can emerge as a consequence of relying on methods for approximating normative probabilistic inference (Fryer, Harms, & Jackson, 2013; Halpern & Pass, 2010; O'Connor, 2006). In contrast, we show that polarization is consistent in some cases with fully normative probabilistic inference. Zimper and Ludwig (2009) made use of nonadditive probability measures, and Dixit and Weibull (2007) showed how repeated voting scenarios can lead to political polarization (Heit & Nicholson, 2010). In contrast, our account relies on standard probability theory and focuses on simple before-and-after judgments like those used in most psychological studies of polariza-
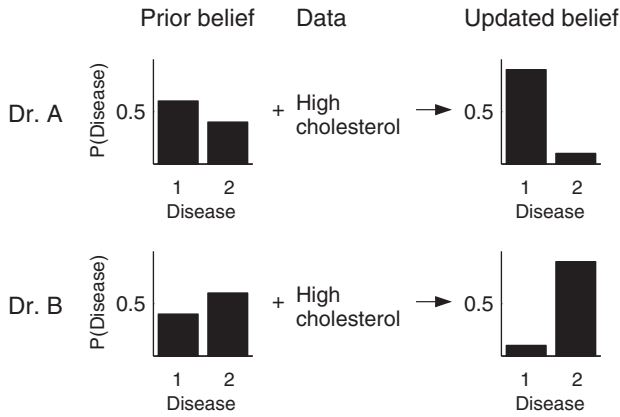
*Figure 1.* An example of belief polarization. Dr. A and Dr. B have opposing prior beliefs about which of two diseases a patient has. They both see the same test result showing that the patient has high cholesterol, and subsequently both become more certain about their initial diagnoses.

tion. We show that even for these kinds of simple judgments, polarization sometimes results from basic probabilistic inference alone.

Our approach contrasts with previous psychological accounts that emphasize the role of motivated reasoning (Klaczynski, 2000; Kunda, 1990) and suggest that polarization results from people interpreting information in a biased manner to favor conclusions that they would like to be true (Dawson, Gilovich, & Regan, 2002; Taber & Lodge, 2006). Previous normative analyses have revealed that apparently irrational phenomena such as confirmation biases (Austerweil & Griffiths, 2011; Navarro & Perfors, 2011; Oaksford & Chater, 1994), reasoning fallacies (Hahn & Oaksford, 2006, 2007; Harris, Hsu, & Madsen, 2012; Oaksford & Hahn, 2004), framing effects (McKenzie, 2004; Sher & McKenzie, 2008), and probability matching (Green, Benson, Kersten, & Schrater, 2010) can be rational in certain contexts. Our analysis suggests that belief polarization is another phenomenon that can be more rational than it appears.

Our primary focus is on belief polarization, but polarization is only one possible outcome that can result when two people update their beliefs. The next section describes the full set of possible outcomes and presents an analysis that reveals the circumstances under which each of these outcomes should be expected. We then use a probabilistic framework to offer normative explanations of the belief polarization observed in several previous studies. Finally, in order to evaluate the extent to which people's behavior is or is not consistent

with our normative account, we describe an experiment that explores how people update their beliefs under different conditions.

## A Formal Analysis of Belief Revision

We treat belief polarization as a special case of *contrary updating*, in which two people update their beliefs in opposite directions after observing the same data (see Figure 2a). *Belief divergence* refers to cases in which the person with the stronger belief in a hypothesis increases the strength of his or her belief and the person with the weaker belief decreases the strength of his or her belief (Figure 2a.i). Divergence therefore includes belief polarization. The opposite of belief divergence is *belief convergence* (Figure 2a.ii), in which the person with the stronger belief decreases the strength of his or her belief and the person with the weaker belief increases the strength of his or her belief.

More formally, consider a situation in which two people observe data $D$ that bear on some hypothesis $H$. Let $P_A(\cdot)$ and $P_B(\cdot)$ be probability distributions that capture the two people's respective beliefs. Contrary updating occurs whenever one person's belief in $H$ increases after observing $D$ and the other person's belief in $H$ decreases after observing $D$, or when

$$[P_A(H \mid D) - P_A(H)][P_B(H \mid D) - P_B(H)] < 0. \quad (1)$$

Contrary updating can be contrasted with *parallel updating* (Figure 2c), in which the two people update their beliefs in the same direction. All situations in which both people change their beliefs after observing data can be unambiguously classified as instances of parallel or contrary updating. It is clear that parallel updating should be the normative outcome in some cases. The conventional wisdom about contrary updating is that divergence is always irrational, but convergence is sometimes rational (Baron, 2008; Gilovich & Griffin, 2010; Munro & Ditto, 1997; Ross & Lepper, 1980).

We show, however, that this conventional wisdom cannot be correct in all circumstances. To see why, suppose that there are two possible hypotheses, $H = 1$ and $H = 2$. Now consider the odds form of $P_A(H \mid D)$:

$$\frac{P_A(H = 1 \mid D)}{P_A(H = 2 \mid D)} = \frac{P_A(D \mid H = 1)}{P_A(D \mid H = 2)} \frac{P_A(H = 1)}{P_A(H = 2)}. \quad (2)$$

A corresponding equation could be written for Person B by replacing $P_A(\cdot)$ with $P_B(\cdot)$. In Equation 2, the value of the likelihood ratio $\dfrac{P_A(D \mid H = 1)}{P_A(D \mid H = 2)}$ determines the direction in which Person A's
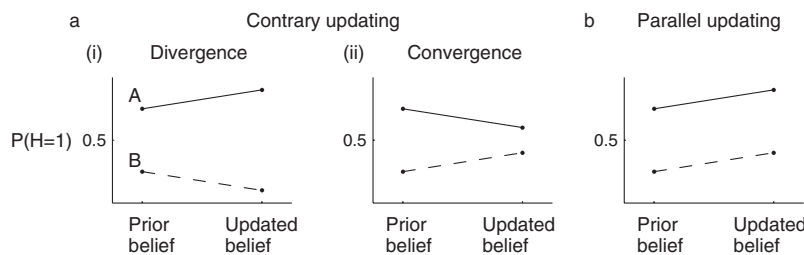


*Figure 2.* Examples of belief revision for two people, A (solid line) and B (dashed line). The two people begin with different beliefs about hypothesis $H$. After observing the same data, their beliefs may (a) move in opposite directions or (b) move in the same direction.

beliefs will change after observing data $D$. If the likelihood ratio is equal to 1, A's beliefs will not change. If the likelihood ratio is greater than 1, A will increase his or her belief that $H = 1$. And if the likelihood ratio is less than 1, A will decrease his or her belief that $H = 1$.

In general, there is no reason to assume that the likelihood ratios for A and B will be equal, or that they will both be greater than or less than 1. Whether A and B update their beliefs in the same direction after observing $D$ depends on the assumptions they each make about the problem. For example, suppose that Smith and Jones started a chess game yesterday and left their unfinished game in the faculty lounge. $D$ represents the current state of the chess board, $H = 1$ indicates that Smith is the stronger player, and $H = 2$ indicates that Jones is the stronger player. Two spectators, Alice and Bob, briefly glance at the chess board, and both agree that white is in the stronger position. However, they arrive at opposite conclusions: Alice concludes that Smith is the stronger player and Bob concludes that Jones is the stronger player. Alice believes that Smith is playing white. Thus, for Alice, the likelihood ratio $\dfrac{P_A(D \mid H = 1)}{P_A(D \mid H = 2)}$ is greater than 1. Bob believes that Smith is playing black. Thus, for Bob, the likelihood ratio $\dfrac{P_B(D \mid H = 1)}{P_B(D \mid H = 2)}$ is less than 1. Even though Alice and Bob are normative reasoners and agree on what qualifies as evidence of strong chess ability, they draw opposite conclusions after seeing the chess board because they made different assumptions (cf. Andreoni & Mylovanov, 2012).

The chess example illustrates that belief divergence can result from normative inference if two people make different assumptions about factors that affect the relationship between hypothesis $H$ and data $D$, such as which person is playing white in the chess game. We now consider what kinds of relationships between variables $H$ and $D$ might or might not give rise to contrary updating. We represent these relationships using Bayesian networks, or Bayes nets for short (Gopnik et al., 2004; Pearl, 2000; Sloman, 2005). In a Bayes net, each variable is represented by a node, and directed edges (arrows) between nodes capture dependence relationships. Note that a Bayes net is only a graphical representation of these probabilistic dependencies between variables. Thus, while the analysis that follows depends on probabilistic inference, it does not depend in any fundamental way on the Bayes net representation. The value of Bayes nets for our purposes is that they offer a convenient way to visualize and conceptualize different types of relationships between variables (Edwards & Fasolo, 2001). As we show, whether belief divergence can result from normative probabilistic inference depends critically on how the variables in a situation are related.

The Bayes net in Figure 3a.i captures a simple relationship in which the data $D$ depend on hypothesis $H$. That is, the probability of observing a given piece of data will depend on which hypothesis is true. We assume throughout that there are exactly two mutually exclusive hypotheses, which means that variable $H$ has only two possible values. For example, $H = 1$ and $H = 2$ might denote whether a patient has Disease 1 or 2, respectively. A Bayes net captures a complete probability distribution over its variables when the network structure is supplemented with a set of conditional probability distributions (CPDs). Figure 3b shows a set of CPDs for Bayes net 3a.i that would produce behavior like that of Dr. A in Figure 1. In this example, $H$ represents the patient's disease and $D$ is a test result, which we assume has only two possible outcomes. The CPD near node $H$ in the figure indicates that Dr. A has a prior belief in favor of Disease 1. The CPD near node $D$ indicates that if the patient has Disease 1, the test is likely to produce Outcome 1. Similarly, if the patient has Disease 2, the test is less likely to produce Outcome 1 and is more likely to produce Outcome 2. Thus, if the test result indicates Outcome 1, Dr. A will become even more certain that the patient has Disease 1.

If two people have different beliefs about the relationships between the variables in a situation, it is clear that they may draw wildly divergent conclusions from the same data (Jaynes, 2003, chapter 5). We therefore focus on cases in which two people agree on the basic structure of a situation. Namely, we assume that both



*Figure 3.* Bayesian networks. (a) (i) A simple Bayes net that captures the relationship between hypothesis $H$ and data $D$. (ii)–(viii) Bayes nets that include an additional variable $V$. Bayes nets in Family 1 can produce only parallel updating. Bayes nets in Family 2 can produce both parallel and contrary updating. (b) A version of Bayes net a.i with conditional probability distributions that produce behavior like that of Dr. A in Figure 1.

people agree on the relevant variables in a situation, and agree on the Bayes net structure and CPDs that capture relationships between these variables. The only allowable difference is that the two people may have different prior beliefs about the values of the root nodes in the Bayes net, where a root node is a node without parents (e.g., node $H$ in Figure 3a.i). Under these assumptions, many Bayes nets are incapable of producing contrary updating. The simple Bayes net in Figure 3a.i is one example. Although the CPDs in Figure 3b can account for the behavior of Dr. A, there is no CPD for the conditional probability $P(D \mid H)$ that can explain the behavior of both doctors, even if they have different prior beliefs, $P_A(H)$ and $P_B(H)$. Because both doctors agree on $P(D \mid H)$, they must also agree that any piece of data $D$ either supports $H = 1$ more strongly or supports $H = 2$ more strongly. Therefore, both divergence and convergence are impossible in this situation.

The Bayes net in Figure 3a.i, however, is too simple to capture the structure of many situations. If a third variable affects the outcome of $D$, belief divergence can be consistent with normative probabilistic inference. For example, the Bayes net in Figure 3a.vi can capture the belief divergence exhibited by the two doctors in Figure 1 if the factor $V$ represents whether the patient has low or high blood sugar, and this factor affects the meaning of the test result $D$. For instance, suppose that a high cholesterol test result is most probable when a patient has Disease 1 and low blood sugar, or when a patient has Disease 2 and high blood sugar. Then two doctors with different prior beliefs about the patient's blood sugar level may draw opposite conclusions about the most probable disease upon seeing the same cholesterol test result $D$.

More generally, as philosophers (Duhem, 1954; Putnam, 1974; Quine, 1951) and psychologists (Cummins, Lubart, Alksnis, & Rist, 1991) have argued, hypotheses are rarely considered in isolation, and inferences about one hypothesis typically depend on additional hypotheses and beliefs. We therefore expand our focus to include the rest of the Bayes nets in Figure 3a, each of which contains one additional variable $V$. Many real-world problems involve more than three variables, but the space of three-node Bayes nets will be sufficient for our purposes. We restrict our attention to cases in which no variables are dependent on $D$, motivated by the idea that the observed data are the final result of a generative process. We also exclude less interesting cases in which the three variables are not all linked in some way (i.e., we consider only *connected* Bayes nets). The remaining Bayes nets

capture cases in which (ii) $V$ is an additional factor that bears on $H$, (iii) $V$ informs the prior probability of $H$, (iv)–(v) $D$ is generated by an intervening variable $V$, (vi) $V$ is an additional generating factor of $D$, (vii) $H$ and $D$ are both effects of $V$, and (viii) $V$ informs both the prior probability of $H$ and the probability of $D$.

In Appendix A, we prove that the Bayes nets in Figure 3a fall into two families. Bayes nets in Family 1 are incapable of producing contrary updating and Bayes nets in Family 2 are capable of producing contrary updating under some circumstances. For Bayes nets in Family 1, all paths from root nodes to $D$ pass through $H$. As a result, even if two people have different prior beliefs about the variables represented by the root nodes, they must make identical inferences about how data $D$ bear on hypothesis $H$. By contrast, the Bayes nets in Family 2 include paths from the root nodes to $D$ that do not pass through $H$, which allows for the possibility that background knowledge can influence how $D$ bears on $H$. The next section demonstrates by example that all three networks in Family 2 can produce belief divergence.

Our analysis undermines the conventional wisdom that belief divergence is always irrational but that convergence and parallel updating are sometimes rational. For example, the Bayes net in Figure 3a.i cannot produce contrary updating. This means that it cannot account for belief divergence, consistent with previous authors' claims that divergence is not rational. However, our analysis shows that the same Bayes net cannot account for belief convergence either, and the same conclusion applies to all of the Bayes nets in Family 1. In contrast, the Bayes nets in Family 2 can account for both convergence and divergence. In other words, if a given network structure predicts that convergence is normative in some cases, it must also predict that divergence is normative in other cases.

## Bayes Net Accounts of Previous Studies of Belief Divergence

The previous section established normative divergence as a theoretical possibility. We now show that it is possible to develop normative accounts of many previous studies of belief polarization. The studies considered in this section are listed in Table 1. The authors of these studies have generally argued that belief divergence emerges as a consequence of processing biases (Liberman & Chaiken, 1992; Lord et al. 1979; McHoskey, 1995; Munro

Table 1
*Previous Studies of Belief Divergence*

| Study | Beliefs about | Evidence provided |
|---|---|---|
| Lord, Ross, and Lepper (1979) | Death penalty | Two conflicting studies |
| Liberman and Chaiken (1992) | Effects of caffeine | Two conflicting studies |
| McHoskey (1995) | JFK assassination | Two opposing theories |
| Munro and Ditto (1997) | Gay people | Two conflicting studies |
| Taber and Lodge (2006) | Multiple issues | Participant-selected arguments |
| Taber, Cann, and Kucsova (2009) | Multiple issues | Two opposing arguments |
| Plous (1991) | Nuclear power safety | Description of an averted catastrophe |
| Batson (1975) | Religion | Story undermining a religious tenet |

*Note.* Each row indicates the subject of the study and the evidence that was provided to participants before measuring their change in beliefs. Studies above the line used mixed evidence and studies below the line used a single piece of evidence.

& Ditto, 1997; Plous, 1991) or motivated reasoning (Taber, Cann, & Kucsova, 2009; Taber & Lodge, 2006). We suggest, however, that the belief divergence in some of these studies may be consistent with normative probabilistic inference. In discussing these studies, we describe Bayes nets that illustrate how all three of the network structures in Family 2 of Figure 3a can produce belief divergence.

## Lord et al. (1979): Beliefs About the Death Penalty

The most widely cited study of belief divergence, conducted by Lord et al. (1979), explored how people update their beliefs about the effectiveness of the death penalty as a crime deterrent after seeing mixed evidence (see also Pomerantz, Chaiken, & Tordesillas, 1995). In this study, supporters and opponents of the death penalty were asked to read about two fictional studies. One study supported the idea that the death penalty is an effective crime deterrent and the other study supported the idea that the death penalty is not an effective crime deterrent. After reading the studies, death penalty supporters strengthened their belief in the effectiveness of the death penalty as a crime deterrent and death penalty opponents weakened their belief. Lord et al. explained the belief divergence as a consequence of an irrational processing bias, but we use Bayes nets from Family 2 in Figure 3a to construct two alternative explanations.

**Data generated by multiple factors.** Our first alternative explanation of the death penalty study is based on two simple assumptions that a participant might have made. The first assumption is that studies—like the ones participants read about—are influenced by research bias, such that researchers tend to arrive at conclusions that are consistent with their own prior beliefs. The second assumption is that one's own beliefs about the effectiveness of the death penalty differ from the consensus opinion among researchers and other experts. This second assumption is similar to the false uniqueness effect, whereby people sometimes expect that their own beliefs are not shared by people in other groups (Mullen, Dovidio, Johnson, & Copper, 1992). We now show that these assumptions can lead to belief divergence through normative probabilistic inference.

These assumptions can be captured using the Bayes net in Figure 4a.i, in which the data $D$ are produced by two factors. Let $H = 1$ correspond to the hypothesis that the death penalty is an effective crime deterrent and $H = 0$ correspond to the hypothesis that it is not. Similarly, let $D = 1$ correspond to a study supporting the idea that the death penalty is an effective crime deterrent (positive evidence) and $D = 0$ correspond to a study supporting the idea that it is not (negative evidence). Finally, let $V = 1$ indicate that the consensus expert opinion supports the effectiveness of the death penalty as a crime deterrent and let $V = 0$ indicate that the consensus expert opinion supports the ineffectiveness of the death penalty as a crime deterrent. The CPD for the $D$ node in Figure 4a.i shows one way that the hypothesis $H$ and the consensus opinion $V$ might jointly shape the outcome of a study. If the consensus opinion ($V$) about the effectiveness of the death penalty ($H$) is correct, there is a high probability that a study will provide support for the true value of $H$. However, if the consensus opinion is incorrect, studies of the death penalty might be influenced by researcher bias. For simplicity, we assume that the two study

outcomes are equally probable when the consensus opinion is incorrect.

The CPDs for the $H$ and $V$ nodes capture the prior beliefs that two study participants, Alice and Bob, might have. For example, Alice initially believes that the death penalty is an effective crime deterrent, $P(H = 1) = 0.8$, but thinks that her belief about the death penalty is different from the consensus expert opinion, $P(V = 1) = 0.2$. Bob's beliefs are the opposite of Alice's. Under these conditions, Figure 4a.ii shows how Alice and Bob should normatively update their beliefs after seeing two conflicting death penalty studies. In this plot, the prior beliefs show $P(H = 1)$ for Alice and Bob. To compute their updated beliefs, we conditioned on the observation of two conflicting pieces of data $D_1 = 0$ and $D_2 = 1$. The updated beliefs in the plot show $P(H = 1 \mid D_1, D_2)$ for Alice and Bob. Alice's prior belief that the death penalty is an effective crime deterrent provides her with a justification for treating the study supporting the opposite conclusion as a spurious result due to researcher bias. Consequently, the two studies combined provide additional support for Alice's prior belief and she becomes more certain. For the same reason, Bob becomes more certain about his belief that the death penalty is not an effective crime deterrent, resulting in belief divergence.

In addition to demonstrating that mixed evidence led to polarization, Lord et al. (1979) also examined how people changed their beliefs as they encountered each piece of evidence. Figure 5a shows how death penalty supporters and opponents in the study changed their beliefs about the effectiveness of the death penalty as a crime deterrent after evaluating each study.[1] The top plot shows participants' average belief change when they were first shown the study providing negative evidence of a crime deterrence effect and the bottom plot shows participants' average belief change when they were first shown the study providing positive evidence of an effect. The overall belief divergence can be observed in both plots by looking at participants' change in beliefs after evaluating both studies: Regardless of the order in which the evidence was presented, death penalty supporters and opponents changed their beliefs in opposite directions. After participants had seen only one of the studies, the two groups changed their beliefs in the same direction, although death penalty supporters changed their beliefs more in response to the positive evidence than to the negative evidence and death penalty opponents changed their beliefs more in response to the negative evidence than to the positive evidence.

A probabilistic approach can also account for the sequences of changes in Figure 5a. The left column of Figure 5b shows the changes in belief predicted by the Bayes net in Figure 4a.i. These predictions were generated by first computing $P(H = 1 \mid D_1)$ and then $P(H = 1 \mid D_1, D_2)$ for each person, where one of $D_1$ and $D_2$ was positive evidence, and the other was negative evidence. Several studies suggest that subjective estimates of uncertainty often correspond to log-odds (Gonzalez & Wu, 1999; Phillips & Edwards, 1966; Tversky & Kahneman, 1992). Therefore the $y$-axis of

---

[1] In the study, after participants read a summary of each death penalty study, they were provided with additional details and critiques of the study. Each participant's cumulative change in belief was recorded once after reading the study summary and again after reading the additional materials. In Figure 5a, we show only the change in belief after reading each study *and* the corresponding additional materials.

a

(i)   H: Death penalty is effective    V: Consensus supports effectiveness    (ii)

|  | P(H=1) |
|---|---|
| Alice | 0.8 |
| Bob | 0.2 |

|  | P(V=1) |
|---|---|
| Alice | 0.2 |
| Bob | 0.8 |

H → D ← V

D: Study supports effectiveness

| H | V | P(D=1\|V,H) |
|---|---|---|
| 0 | 0 | 0.1 |
| 0 | 1 | 0.5 |
| 1 | 0 | 0.5 |
| 1 | 1 | 0.9 |

(ii) plot: P(H=1) 0.5; A, B; Prior belief → Updated belief

b

(i)   H: Death penalty is effective    V: Deterring strength of death penalty    (ii)

| V | P(H=1\|V) |
|---|---|
| 1 | 0 |
| 2 | 0 |
| 3 | 1 |
| 4 | 1 |

|  | P(V=1) | P(V=2) | P(V=3) | P(V=4) |
|---|---|---|---|---|
| Alice | 0.1 | 0.1 | 0.2 | 0.6 |
| Bob | 0.6 | 0.2 | 0.1 | 0.1 |

H ← V → D

D: Deterring strength supported by study

| V | P(D=1\|V) | P(D=2\|V) | P(D=3\|V) | P(D=4\|V) |
|---|---|---|---|---|
| 1 | 0.7 | 0.1 | 0.1 | 0.1 |
| 2 | 0.1 | 0.7 | 0.1 | 0.1 |
| 3 | 0.1 | 0.1 | 0.7 | 0.1 |
| 4 | 0.1 | 0.1 | 0.1 | 0.7 |

(ii) plot: P(H=1) 0.5; A, B; Prior belief → Updated belief

c

(i)   H: Jesus is son of God    V: Christian world view    (ii)

| V | P(H=1\|V) |
|---|---|
| 0 | 0.1 |
| 1 | 0.9 |

|  | P(V=1) |
|---|---|
| Alice | 0.9 |
| Bob | 0.1 |

H ← V, H → D ← V

D: Christians are persecuted

| H | V | P(D=1\|V,H) |
|---|---|---|
| 0 | 0 | 0.4 |
| 0 | 1 | 0.4 |
| 1 | 0 | 0.01 |
| 1 | 1 | 0.6 |

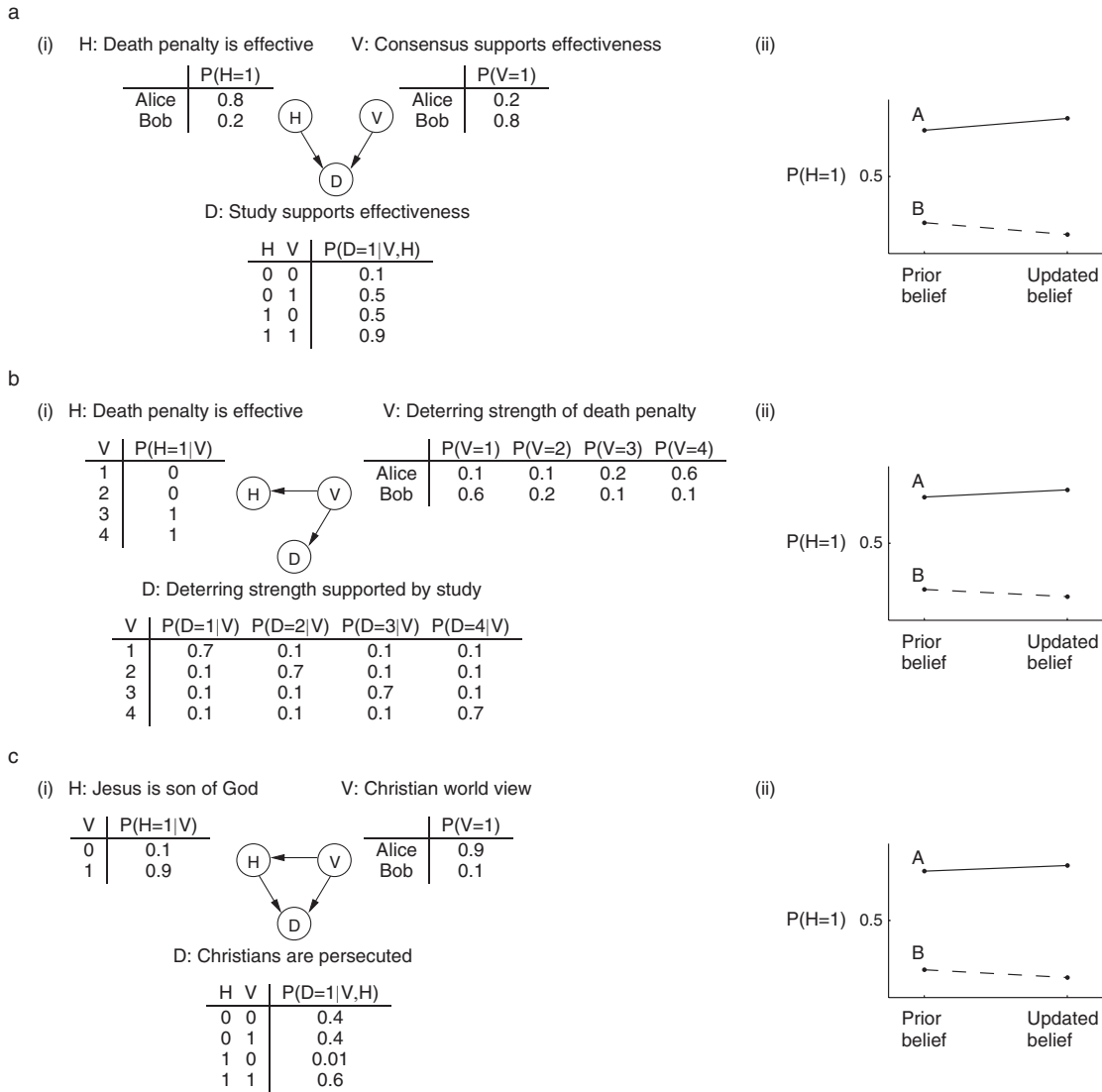(ii) plot: P(H=1) 0.5; A, B; Prior belief → Updated belief

*Figure 4.* Example Bayes nets that produce the normative belief divergence shown on the right. The labels above each conditional probability distribution indicate the meanings of the corresponding variables. (a–b) Two examples that may explain belief divergence reported by Lord et al. (1979). (c) An example that may explain the belief divergence reported by Batson (1975).

each plot in Figure 5b shows differences on a log-odds (or logit) scale[2] As the figure shows, the Bayes net model captures most of the qualitative effects evident in the data. Consider first the plot in the top row. When a study providing negative evidence is observed first, the model correctly predicts that both supporters and opponents will weaken their belief in the effectiveness of the death penalty, and that opponents will weaken their beliefs to a greater extent. When the second study providing positive evidence is observed, the model predicts that supporters and opponents will both strengthen their beliefs in the effectiveness of the death penalty, but that supporters alone will conclude the experiment with a net gain in the strength of these beliefs.

The model predictions in the bottom row are the mirror image of the predictions in the top row. We expect that any normative account should make symmetric predictions in these two cases provided that the negative evidence supports the ineffectiveness of the death penalty to the same degree that the positive evidence supports the effectiveness of the death penalty. The data in the bottom row of Figure 5a, however, are not quite equivalent to the data in the top row: After participants observed the negative evidence, death penalty opponents weakened their beliefs as expected, but contrary to the model's predictions, supporters slightly strengthened their beliefs. Lord et al. (1979), however, did not report measures of variance, which makes it difficult to assess the

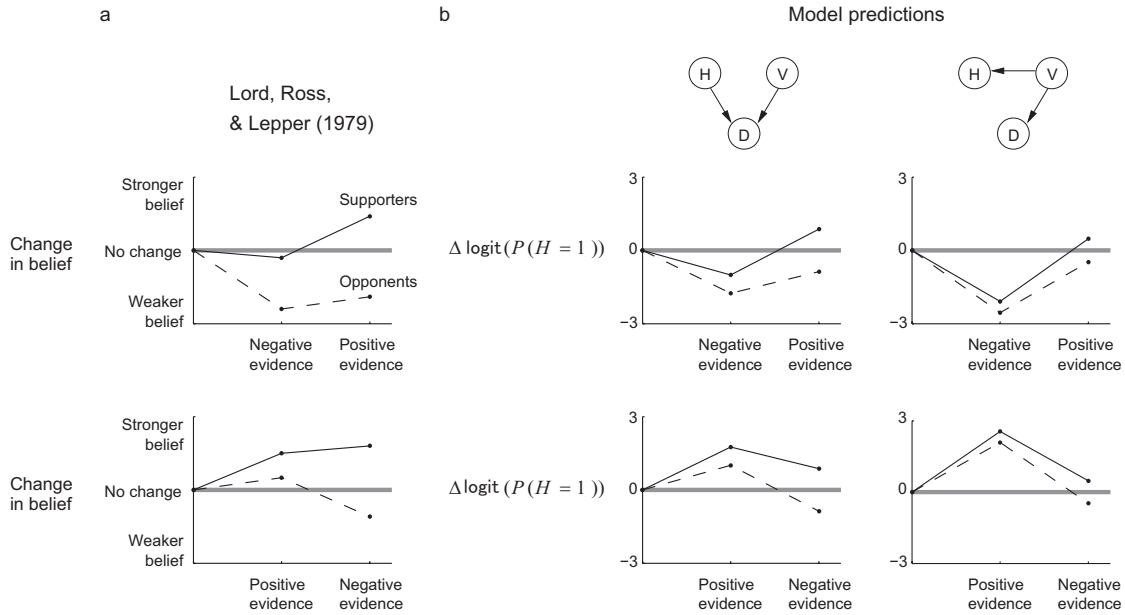$$^2 \, \mathrm{logit}(P(H = 1)) = \log\left(\frac{P(H = 1)}{1 - P(H = 1)}\right)$$

*Figure 5.* Data and model predictions for belief change over time. All plots show cumulative change in belief about the effectiveness of the death penalty among death penalty supporters (solid line) and opponents (dashed line). The thick gray line indicates no belief change. The top row shows belief change when observing evidence supporting the ineffectiveness of the death penalty (negative evidence), followed by evidence supporting the effectiveness of the death penalty (positive evidence). The bottom row shows belief change when the evidence is observed in the reverse order. (a) Data reproduced from Lord et al. (1979). Participants indicated their change in belief using a scale ranging from −8 to +8. (b) Predicted belief change for the two Bayes nets in Figures 4b.i and 4c.i.

significance of this discrepancy between the model predictions and the data.

**Hypothesis and data both generated by another factor.** We now consider a second qualitatively different way in which a probabilistic approach can account for the divergence observed in the death penalty study. Here we suppose that participants' responses are based on beliefs about the strength of the effect that the death penalty has on crime deterrence and that belief divergence emerges as a consequence of mapping an ordinal variable (the strength of the effect) onto the binary variable used in the study (whether the death penalty is an effective deterrent).

Figure 4b.i shows a Bayes net that captures this second approach to the death penalty study. Variable $V$ represents the strength of the effect that the death penalty has on crime deterrence, and we assume that $V$ ranges from 1 (*very weak effect*) to 4 (*very strong effect*). $H$ represents the binary variable that participants were asked to make judgments about, with $H = 1$ once again corresponding to the hypothesis that the death penalty is an effective crime deterrent and $H = 0$ corresponding to the hypothesis that it is not. We assume that whether the death penalty is judged to be an effective crime deterrent is determined by the strength of the effect of the death penalty on crime deterrence. The CPD for the $H$ node indicates that the death penalty is deemed effective if the value of $V$ is on the "strong effect" half of the range and ineffective if the value of $V$ is on the "weak effect" half of the range. We also assume that a piece of data $D$ may range from support for a very weak to a very strong effect of the death penalty on crime deterrence. We assume that $D$ ranges from 1 (*evidence of a very weak*

*effect*) to 4 (*evidence of a very strong effect*). The CPD for the $D$ node in Figure 4b.i indicates that there is a relatively high probability that any given study will provide evidence for the true strength of the effect $V$ of the death penalty, but there is a probability that it will provide false support for a different conclusion.

The CPD for the $V$ node captures prior beliefs that Alice and Bob might have. Alice believes that it is most probable that the death penalty has a very strong or moderately strong effect on crime deterrence and Bob believes the opposite. Under these conditions, Figure 4b.ii shows how Alice and Bob should normatively update their beliefs after seeing one study supporting a very weak effect ($D = 1$) and one study supporting a very strong effect ($D = 4$). Like in the previous example, Alice's prior belief that a weak effect is very unlikely provides her with a justification for treating the study supporting a very weak effect as noise, making the study supporting the very strong effect more persuasive. Alice therefore becomes even more certain about her belief, as does Bob, resulting in belief divergence.

We also used this Bayes net model to account for the data in Figure 5a showing participants' belief change over time. The model predictions are shown on the right of Figure 5b. The model correctly predicts that supporters' beliefs will change more in response to the positive than the negative evidence and that opponents' beliefs will change more in response to the negative than the positive evidence.

Our two Bayes net accounts of the death penalty study do not imply that participants diverged in this study for normative rea-

sons. Both accounts rely on assumptions that go beyond the materials provided by the experimenters, and these assumptions may or may not be accurate. We do not claim that either account is the correct explanation for the study results. We propose only that these accounts are plausible explanations that cannot be ruled out a priori. As a result, the data summarized in Figure 5a do not provide definitive evidence of irrational behavior, and additional evidence is needed to demonstrate that the experimental paradigm used by Lord et al. (1979) produces divergence for reasons that are incompatible with normative probabilistic inference.

**Similar studies.** Several other studies of belief divergence have used experimental paradigms that are conceptually similar to the one used in the death penalty study (Liberman & Chaiken, 1992; McHoskey, 1995; Munro & Ditto, 1997; Taber et al. 2009). Participants in these studies were asked to read two opposing studies or arguments, and their resulting changes in belief were measured. Assumptions like those captured by the two Bayes nets for the death penalty study could therefore be applied to these studies.

Liberman and Chaiken (1992) asked non-coffee drinkers and heavy coffee drinkers to read summaries of two studies. One study supported a link between coffee drinking and a disease and the second study did not support such a link. After reading the studies, non-coffee drinkers increased their beliefs that drinking coffee caused the disease but heavy coffee drinkers did not change their beliefs. As with the death penalty study, it is difficult to know what assumptions participants in the coffee study made. Because the procedure of the coffee study is similar to the one in the death penalty study, participants may have made assumptions that correspond to those captured by the Bayes nets in Figures 4a.i and 4b.i.

Taber et al. (2009) explored participants' beliefs and attitudes regarding eight different issues, such as marijuana legalization and tuition increases. Across all issues, the authors found that participants with strong prior beliefs were likely to show belief divergence, and participants with weak prior beliefs were not. A similar result was reported by McHoskey (1995), who asked supporters of the official account of the John F. Kennedy assassination and supporters of an alternative conspiracy account to read a summary of arguments for each account. Those with strong prior beliefs diverged and those with weak prior beliefs did not. The Bayes nets presented in this section cannot account for the fact that only the participants in these studies with strong prior beliefs diverged. However, the results of these studies cannot rule out the possibility that the participants that diverged made assumptions similar to those captured by the Bayes net models we developed for the death penalty study.

Munro and Ditto (1997) used a paradigm in which participants received mixed evidence in a study of beliefs and attitudes about stereotypes of gay people. Participants in the study read one study supporting common stereotypes and another study refuting these stereotypes. Participants' beliefs about propositions regarding the behavior of gay people did not diverge, but participants' general attitudes toward gay people did diverge. Participants who reported a low acceptance of homosexuality became less accepting after reading the studies, and participants who reported a high acceptance of homosexuality became more accepting. The Bayes nets described so far include variables that represent propositions (e.g., "the death penalty deters crime") rather than attitudes (e.g., "I am in favor of the death penalty"). It is possible, however, that similar Bayes nets can be used to explain cases where attitudes diverge.

## Batson (1975): Religious Beliefs

The belief divergence studies discussed so far have all involved mixed evidence. We now consider a study in which two groups of people diverged after observing the same single piece of evidence. In this study, Batson (1975) asked participants with strong Christian beliefs and participants with weak Christian beliefs to read a story describing how church leaders had conspired to cover up new evidence that undermined the idea that Jesus is the son of God. After reading this story, the less religious participants became less certain about their religious beliefs and the more religious participants became more certain about their beliefs, resulting in belief divergence (see Figure 6). Batson hypothesized that the more religious participants may have defensively overcompensated in adjusting their beliefs as a response to the threatening information. We provide an alternative normative explanation based on the Bayes net in Figure 3a.viii.

**Hypothesis and data both informed by another factor.** Suppose that the different beliefs of the two groups not only influenced their judgments about whether Jesus is the son of God but also influenced their expectations about what the data would mean if he were. This idea can be captured using the Bayes net in Figure 4c.i, in which a third factor informs both beliefs about $H$ and about how $D$ is generated. Specifically, suppose that the additional factor $V$ represents a worldview. For instance, someone with a "Christian" worldview ($V = 1$) believes that Jesus is probably the son of God, and that followers of Jesus are likely to have their faith challenged by others. Someone with a "secular" worldview ($V = 0$) believes that Jesus is probably not the son of God, but that if he were, his followers would be unlikely to encounter challenges to their faith. Let $H = 1$ correspond to the hypothesis that Jesus is the son of God and $H = 0$ correspond to the hypothesis that he is not. Let $D = 1$ correspond to an observation that Christians' faith is frequently challenged. The story in the study would be a significant example of something that should challenge Christians' faith. Let $D = 0$ correspond to an observation that Christians' faith is not frequently challenged.

In this example, one's worldview $V$ influences one's beliefs about the hypothesis $H$ as well as one's interpretation of the data $D$. The CPD for the $H$ node in Figure 4c.i indicates that someone with a "Christian" worldview will place a high probability on the hypothesis that Jesus is the son of God and someone with a "secular" worldview will place a low probability on the hypothesis that Jesus is the son of God. The CPD
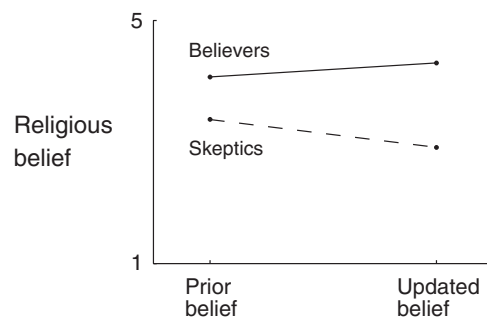


*Figure 6.* Data reproduced from Batson (1975).

for the *D* node indicates how *H* and *V* jointly influence how someone will interpret a piece of data. The exact probabilities were chosen to reflect the fact that, regardless of worldview, people will agree on a base rate of challenges to one's faith if Jesus is not the son of God but that more frequent challenges are expected under the "Christian" worldview than the "secular" worldview.

The CPD for the *V* node captures prior beliefs that Alice and Bob might have. Alice places a high probability on the "Christian" worldview and Bob places a high probability on the "secular" worldview. Under these conditions, Figure 4c.ii shows how Alice and Bob should normatively update their beliefs after seeing evidence that followers of Jesus have had their faith challenged (*D* = 1). Because of Alice's and Bob's different worldviews, they disagree about whether this observation provides support for or against the hypothesis that Jesus is the son of God, and they diverge as a result.

**Similar study.** In a study of beliefs about nuclear power safety, Plous (1991) used an experimental paradigm broadly similar to the one used in the religious belief study. Like the religious belief study and unlike the death penalty study and related studies, Plous observed belief divergence in an experiment in which participants from opposing groups were provided with the same single piece of evidence. In the experiment, supporters and opponents of nuclear power read identical descriptions of an actual nuclear power plant breakdown that was safely contained. After reading these descriptions, nuclear power supporters became more certain that nuclear power is safe and nuclear power opponents became more certain that nuclear power is unsafe.

Because of the procedural similarity of the nuclear power study to the religious belief study, it may be possible to develop a similar normative account of the observed belief divergence based on the idea that the two groups of people were reasoning under different worldviews. For instance, nuclear power supporters may have had a "fault-tolerant" worldview, in which breakdowns are inevitable but are likely to be safely contained. A fault-tolerant worldview would lead someone to believe that nuclear power is safe. Nuclear power opponents may have had a "fault-free" worldview, in which all breakdowns are viewed as dangerous. A fault-free worldview would lead someone to believe that nuclear power is unsafe. After learning about a safely contained breakdown, someone with a fault-tolerant worldview would treat this as an likely outcome in a safe plant and someone with a fault-free worldview would treat any breakdown as a likely outcome in an unsafe plant, causing their beliefs to diverge.

## Summary

Each of our alternative explanations represents one possible characterization of a previous study. We do not claim that these characterizations are necessarily accurate. Rather, we claim that they are plausible accounts of what happened in previous studies that cannot be ruled out on the basis of the evidence that the studies provide. Consequently, the results of these studies cannot be taken as definitive evidence of irrational behavior.

There is one study in Table 1 that we have not discussed. In this study (Taber & Lodge, 2006), participants were allowed to choose

which arguments were presented to them, and therefore, people with opposing beliefs did not necessarily see exactly the same data. Researchers have developed normative models that can select which pieces of data to examine—for example, models that identify the observations that are most likely to distinguish between the current working hypotheses (Austerweil & Griffiths, 2011; Navarro & Perfors, 2011; Oaksford & Chater, 1994). It may be possible to apply these models to the Taber and Lodge (2006) study, but in this article we have chosen to focus on cases in which two people observe the same data.
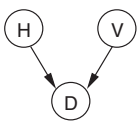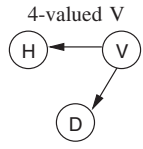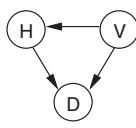
### How Common Is Normative Belief Divergence?

The previous section demonstrated that all three of the Bayes net structures in Family 2 of Figure 3a can produce belief divergence. It is possible, however, that divergence is exceedingly rare within the Bayes nets of Family 2, and that the examples in Figure 4 are unusual special cases that depend on carefully selected CPDs. To examine this possibility, we ran simulations that explored the space of all possible CPDs for the three Bayes net structures in Family 2.

We ran two simulations for each Bayes net structure. In one simulation, we sampled the priors and each row of each CPD from a symmetric Beta distribution with parameter 0.1, resulting in probabilities highly biased toward 0 and 1. In another simulation, we sampled all probabilities from a uniform distribution. In each trial, we generated a single set of CPDs and then generated two different prior distributions for each root node in the Bayes net to simulate two people, consistent with our assumption that two people may have different priors but must agree on the conditional probabilities. We carried out 20,000 trials in each simulation. We counted trials as instances of divergence only if $|P(H = 1 | D = 1) - P(H = 1)| > 10^{-5}$ for both people.

The results of these simulations are shown in Table 2. Because one of our Bayes net accounts of the death penalty study used a 4-valued *V* variable (Figure 4b.i), we also used a 4-valued *V* variable in our simulations for the corresponding network structure in the second

Table 2

*Proportion of Simulation Trials That Produced Belief Divergence Using the Specified Bayes Net Structure (Column) and Probability Distributions (Row)*



| Probability distribution | Bayes net structure | | |
|---|---|---|---|
| | | 4-valued *V* | |
| Biased | 4.8% | 11.7% | 6.4% |
| Uniform | 9.1% | 10.0% | 8.0% |

*Note.* The prior and conditional probabilities for the simulation results shown in the first and third columns were sampled from a beta(0.1, 0.1) distribution (biased) or a beta(1, 1) distribution (uniform). The prior and conditional probabilities for the simulation results shown in the second column were sampled from a Dirichlet([0.1, 0.1, 0.1, 0.1]) distribution (biased) or a Dirichlet([1, 1, 1, 1]) distribution (uniform).

column of the table.[3] In all cases, belief divergence was produced in a non-negligible proportion of trials. These results suggest that divergence is not uncommon within the space of Family 2 Bayes nets. Because the frequency of divergence was similar regardless of whether the CPDs were sampled from a biased or a uniform distribution, our results also suggest that belief divergence does not depend critically on particular settings of the CPDs.

Given that belief divergence does not seem rare in the space of all Bayes nets, it is natural to ask whether cases of normative divergence are regularly encountered in the real world. One approach to this question would be to compile a large database of networks that capture everyday belief revision problems, and to determine what proportion of these networks lead to normative divergence. Based on our simulations, we predict that divergence is relatively likely to arise in situations in which there are more than two variables of interest. Here, however, we turn to a second question motivated by our analyses and ask whether normative belief divergence is consistent with human behavior.

## Experiment

Earlier, we presented Bayes net models that are able to account for the belief divergence observed in several previous empirical studies. It is difficult, however, to determine whether the divergence observed in these studies was genuinely normative without knowing the assumptions and prior beliefs that participants brought to the tasks in question. To evaluate our Bayes net account, we therefore developed a new task that provided us with more control over what participants assumed about the structure of the situation. We used this task to test a specific prediction of our normative analysis: that belief divergence can be made more or less likely by manipulating people's prior beliefs.

Like the death penalty study by Lord et al. (1979), our task explores how participants respond to mixed evidence. We used a medical diagnosis scenario captured by the Bayes net in Figure 7. Suppose that a patient has one of four different diseases, represented by $V$, but that the treatment depends only on what *class* of disease $H$ the patient has. Two of the diseases (L1 and L2) are "allozedic" diseases and the other two (Y1 and Y2) are "hypozedic" diseases. There is a test $D$ for the four diseases that is reasonably accurate but that sometimes provides spurious results. Dr. A and Dr. B have different prior beliefs about the four individual diseases (CPD for $V$), but they both agree about the class of each disease (CPD for $H$) and the accuracy of the test (CPD for $D$). Consistent with their priors, Dr. A initially believes that the disease is probably hypozedic, and Dr. B initially believes that the disease is probably allozedic. After seeing one test that indicates Disease L1 and another test that indicates Disease Y1, both doctors strengthen their initial judgments about the class of the disease. To understand this intuitively, consider Dr. A's perspective. The two test results cannot both be accurate, and Dr. A believes Disease Y1 is much more likely than Disease L1. Thus, Dr. A has reason to treat the test result indicating Disease L1 as spurious, leaving only support for Disease Y1, consistent with Dr. A's prior beliefs. Dr. B's perspective is exactly opposite. Note that this scenario is very similar to the death penalty scenario captured by the Bayes net in Figure 4b.

Our normative account predicts that any pattern of behavior in Figure 2 can be produced by pairing judgments from doctors with different prior beliefs. For example, consider four doctors with prior beliefs induced by the charts in Figure 8. Each chart shows the number of patients who previously had each disease. All four doctors observe a test that indicates L1 and a test that indicates Y1, and they all update their beliefs as predicted by our normative account. Pairing doctors (a) and (b) produces divergence, pairing doctors (c) and (d) produces convergence, and pairing doctors (a) and (d) produces parallel movement. Note that the charts in the left and right columns of Figure 8 are identical except that the "hypozedic" and "allozedic" labels have been swapped. Our experiment therefore focused on the two cases in the left column of the figure and used these cases to explore whether the same test results could lead people to update their beliefs in opposite directions.

We provided participants with information that our normative account predicted should make them more or less certain about their initial beliefs, or should have no effect on their initial beliefs. This resulted in three conditions, which we refer to as the *polarization* (more certain), *moderation* (less certain), and *control* (no effect) conditions, respectively. Although our normative account predicts that the three conditions should lead to different inferences, there are three psychologically plausible responses to the contradictory evidence provided in each condition. Participants might see support for two diseases in different classes and therefore become less certain about which class of disease the patient had, they might discount the less likely test result and become more certain, or they might view the two tests as canceling one another out and therefore remain equally certain. All three approaches seem intuitively natural, and we therefore expected that some participants would adopt each approach regardless of their prior beliefs. The main goal of our experiment was to see whether the relative proportion of participants adopting each approach changed, depending on their induced prior beliefs about the four diseases.

### Bayes Net Model Predictions

Predictions of our normative account were based on a version of the Bayes net in Figure 7. The prior distributions for $V$ were derived from the charts in Figure 8. The distributions we used were more skewed than the ones in Figure 7 in order to produce as strong of an effect as possible. We assumed that the most likely result of each test was the true disease and that all other diseases were equally likely to be spurious results, as shown in the CPD for $P(D \mid V)$ in Figure 7. Given this assumption, our probabilistic analysis predicts that the strength of participants' beliefs should increase, decrease, and remain unchanged in the polarization, moderation, and control conditions, respectively. Appendix B shows that these predictions hold regardless of the exact numbers in the CPD for $P(D \mid V)$ and explains that these predictions also hold if the distributions in Figure 8 are distorted according to a weighting function like that used by prospect theory (Tversky & Kahneman, 1992).

---

[3] Jern, Chang, and Kemp (2009) showed that contrary updating is impossible for a Bayes net with this network structure when $V$ is binary.

H: Disease class

| V | P(H=Allozedic|V) |
|---|---|
| L1 | 1 |
| L2 | 1 |
| Y2 | 0 |
| Y1 | 0 |

V: Disease

| | P(V=L1) | P(V=L2) | P(V=Y2) | P(V=Y1) |
|---|---|---|---|---|
| Dr. A | 0.05 | 0.30 | 0.25 | 0.40 |
| Dr. B | 0.40 | 0.25 | 0.30 | 0.05 |

D: Test result

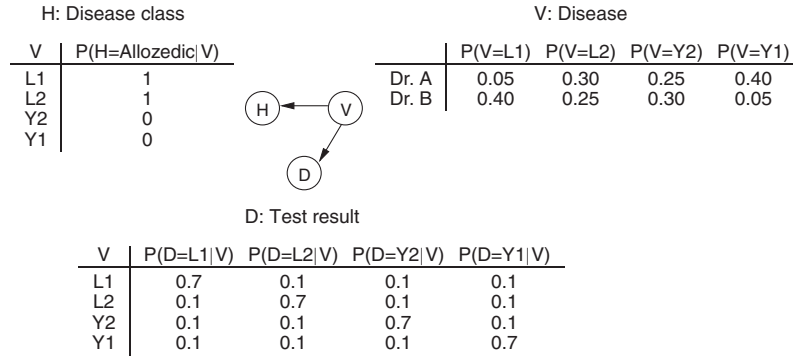| V | P(D=L1|V) | P(D=L2|V) | P(D=Y2|V) | P(D=Y1|V) |
|---|---|---|---|---|
| L1 | 0.7 | 0.1 | 0.1 | 0.1 |
| L2 | 0.1 | 0.7 | 0.1 | 0.1 |
| Y2 | 0.1 | 0.1 | 0.7 | 0.1 |
| Y1 | 0.1 | 0.1 | 0.1 | 0.7 |

*Figure 7.* A version of Bayes net a.viii that captures the assumptions of the medical diagnosis scenario in the experiment. The labels above each conditional probability distribution indicate the meanings of the corresponding variables. L1 and L2 = allozedic diseases; Y1 and Y2 = hypozedic diseases.

## Method

**Participants.** Four hundred seventeen participants were recruited online through Amazon Mechanical Turk. They were paid for their participation.

**Design and materials.** We used a 3 × 2 design. One factor was the information provided to participants, resulting in the polarization, moderation, and control conditions. The other factor



*Figure 8.* Prior distributions that produce different patterns of belief revision according to our normative account. Each distribution is represented as a frequency distribution and shows the number of people with the patient's symptoms, out of 1,000, who previously had each of the diseases. Diseases L1 and L2 are allozedic diseases, and Y1 and Y2 are hypozedic diseases. The labels above each chart indicate the initial diagnosis and the predicted change in belief after observing a test that indicates L1 and a test that indicates Y1. The line plots illustrate the predicted belief revision behavior, where the *y*-axis ranges from "definitely allozedic" to "definitely hypozedic." Note that the charts in the left and right columns are identical except that the "hypozedic" and "allozedic" labels have been swapped.

was whether participants were asked a forced choice question about which test result was more likely to be accurate. The purpose of the forced choice manipulation was to see whether people would be more inclined to change their beliefs when their attention was drawn to the fact that both tests could not be accurate.

All participants initially saw a chart like those in Figure 8 showing the frequency of occurrence for each disease. Although Figures 8a and 8c show different charts, we adjusted the moderation condition so that participants in all three conditions initially saw the chart in Figure 8a, which made the control condition equally similar to each of the other two conditions.

Participants then saw two contradictory test results. In the moderation condition, one test result indicated that the patient most likely had Disease L2 and one test result indicated that the patient most likely had Disease Y2. This manipulation is equivalent to using the chart in Figure 8c with test results indicating Diseases L1 and Y1. In both cases, participants see a test result indicating the more common allozedic disease and the less common hypozedic disease. In the polarization condition, one test result indicated that the patient most likely had Disease L1 and one test result indicated that the patient most likely had Disease Y1. In the control condition, the test results did not distinguish between the individual diseases: one test result indicated that the patient most likely had an allozedic disease and one test result indicated that the patient most likely had a hypozedic disease.

**Procedure.** Participants were randomly assigned to conditions. For participants who were asked the forced choice question, 54 were in the polarization condition, 48 were in the moderation condition, and 53 were in the control condition. For participants who were not asked the forced choice question, 87 were in the polarization condition, 89 were in the moderation condition, and 86 were in the control condition.

The task consisted of two screens, one for each belief judgment. On the first screen, participants saw the chart depicting the frequency of occurrence of each disease and were asked, "Do you think the patient has an allozedic or a hypozedic disease, and how confident are you?" They made their judgments on a scale from −100 (*Absolutely certain that the patient has an allozedic disease*) to +100 (*Absolutely certain that the patient has a hypozedic disease*). The charts always favored the disease class on the positive end of the scale because a pilot study suggested that
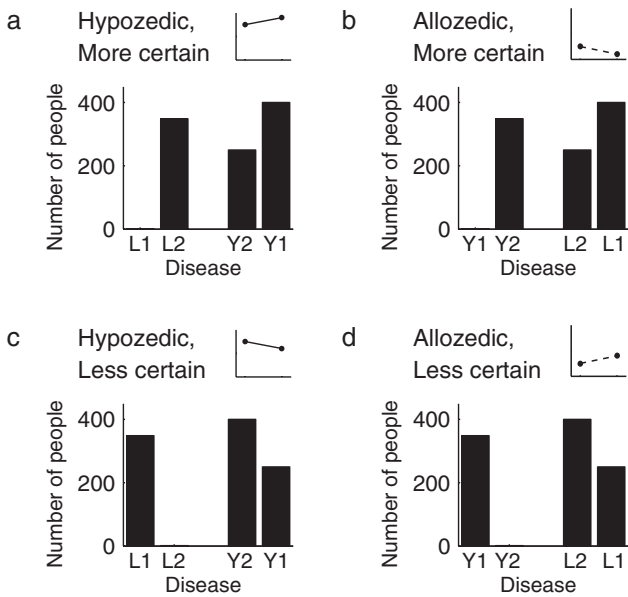
people had difficulty thinking about the negative end of the scale. Whether the more likely disease class was allozedic or hypozedic was randomized across participants.

On the second screen, participants saw the same chart and two test results, presented in random order, along with their initial judgments. They were again asked which class of disease was more likely. Before making their second judgment, participants who received the forced choice question were told "The tests can't both be accurate" and were asked to choose whether the first or second test was more likely to be an accurate result.

## Results and Discussion

Participants were classified as less certain, equally certain, or more certain about their initial judgments of the patient's disease class after seeing the test results. The proportions of participants in these three categories are shown for each condition in Figure 9. Results from conditions with the forced choice question are shown in Figure 9a and results from conditions without the forced choice question are shown in Figure 9b. The confidence intervals in the figure were estimated using the statistical bootstrap technique with 100,000 samples. Responses from 12 participants were eliminated because their initial judgments were on the opposite end of the scale than should have been suggested by the charts in Figure 8, suggesting that they misunderstood some aspect of the task.

Our primary question was whether the relative proportions of the three possible responses changed as a function of participants' prior beliefs. The control conditions provide a baseline for participants' responses when they are provided with contradictory information. Thus, the critical comparisons are between the control conditions and the moderation and polarization conditions. In the control condition, when participants were asked the forced choice question, 39% became less certain, 39% remained equally certain, and 22% become more certain. When participants were not asked the forced choice question, 38% became less certain, 42% remained equally certain, and 20% became more certain.

The plots in the leftmost and rightmost columns of Figure 9 show the comparisons between the control conditions and the other two conditions. When participants were asked the forced choice question, those in the moderation condition were significantly more likely than those in the control condition to become less certain ($z = 2.34$, $p = .010$, one-tailed $z$-test for proportions). Participants in the polarization condition were significantly more likely than those in the control condition to become more certain ($z = 2.33$, $p = .010$, one-tailed). Similarly, when participants were not asked the forced choice question, those in the moderation condition were significantly more likely than those in the control condition to become less certain ($z = 1.87$, $p = .031$, one-tailed), and those in the polarization condition were significantly more
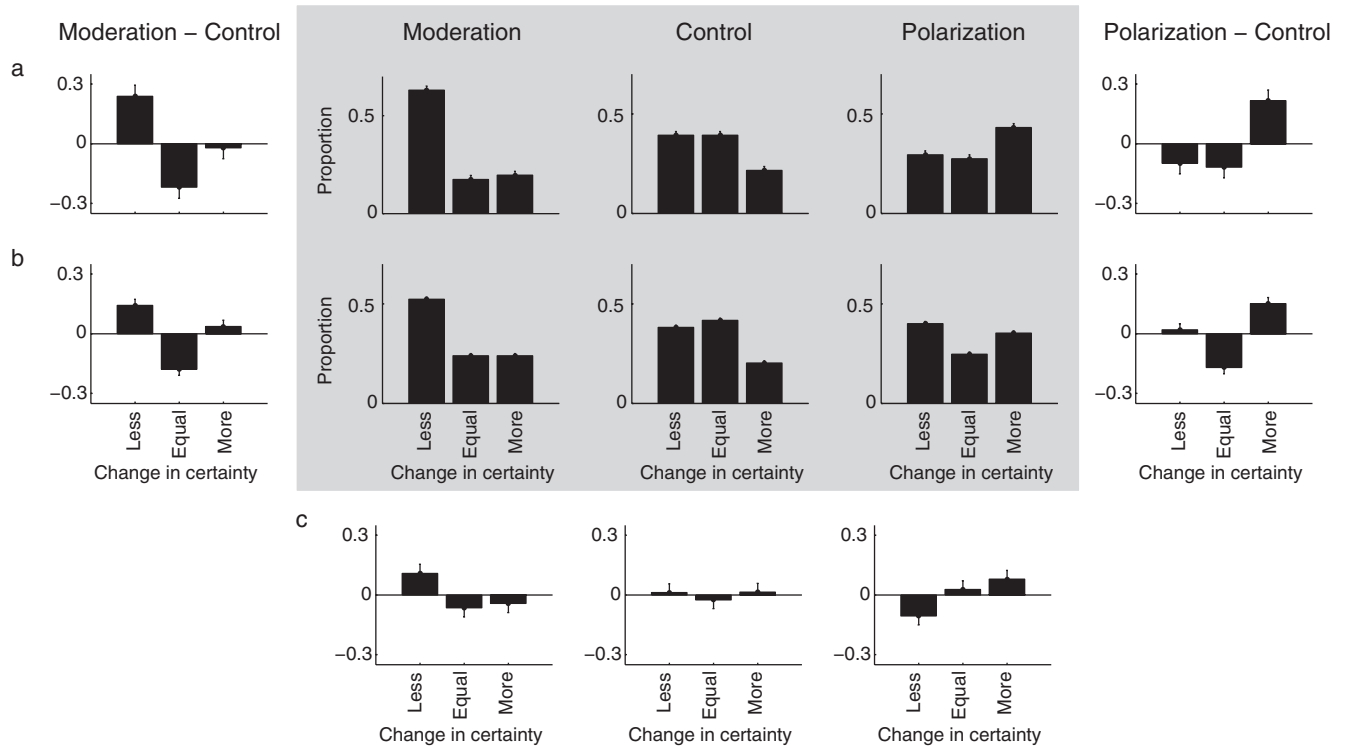


*Figure 9.* Experimental data from (a) conditions with the forced choice question and (b) conditions without the forced choice question. Each plot in the central gray region shows results from one condition in the 3 × 2 design. The three bars in each plot indicate the proportion of participants who became less certain, remained equally certain, or became more certain about their initial beliefs. The plots in the leftmost and rightmost columns outside the gray region show the differences between the moderation and control conditions and between the polarization and control conditions, respectively. (c) Differences between the results for participants who answered the forced choice question and those who did not. In all plots, error bars indicate 95% confidence intervals for the means.

likely than those in the control condition to become more certain ($z = 2.18$, $p = .015$, one-tailed). These results support the idea that people tend to update their beliefs in the direction predicted by our normative account.

Figure 9c shows the effects of the forced choice question on each condition. The forced choice question tended to push participants in the predicted directions, increasing the number of less certain responses in the moderation conditions and the number of more certain responses in the polarization conditions. These effects, however, were not significant ($p > .10$ in both cases). As the figure shows, the forced choice question had almost no effect on the control conditions.

The strongest evidence for normative belief revision in our task would be a set of results in which the most common response in every condition is the normative response. This pattern holds for the moderation conditions, where the most common response was to become less certain, both with (63%) and without (52%) the forced choice question ($p < .001$).[4] In the polarization condition with the forced choice question, the most common response was to become more certain (43%), although this result was not significant ($p = .12$). In the polarization condition without the forced choice question, the most common response (40%) was to become less certain. In the control condition without the forced choice question, the most common response was to remain equally certain (42%), but this result was not statistically significant ($p = .36$). In the control condition with the forced choice question, remaining equally certain and becoming less certain were equally common (39% each).

Overall, our data provide qualified support for our normative account of belief revision. In every condition, some people became less certain about their initial beliefs, some people became more certain, and some people remained equally certain. Many people's responses were therefore inconsistent with our normative account. Critically, however, the proportions of people giving each response changed, depending on their induced prior beliefs, and these changes agreed with the predictions of our normative analysis. In particular, people were more likely to polarize when polarization was the normative response. If people consistently look for evidence supportive of their initial beliefs, our participants could have found it in all conditions. Similarly, if people consistently treat contradictory information as grounds for moderating their beliefs, our participants could have done so in all conditions. Instead, many people appear to have considered the data more carefully with respect to their prior beliefs, consistent with a normative account.

## General Discussion

We presented a normative probabilistic analysis to support the claim that belief polarization should not always be taken as evidence of irrationality. We illustrated this claim by applying our approach to several previous studies of belief divergence and arguing that the results of these studies may be consistent with normative probabilistic inference. To test whether people's inferences are consistent with our normative account, we conducted an experiment using a task in which participants updated their beliefs after seeing two pieces of contradictory evidence. The results suggested that manipulating prior beliefs caused some, but not all, participants to adjust their inferences as predicted by our normative account.

## Biased Evaluation of Evidence

The empirical case for belief divergence rests primarily on measures of belief change, and our discussion of previous studies therefore focused on these measures. Previous studies of divergence, however, often include other kinds of measures. For example, researchers often ask participants to rate the quality of the evidence provided and to supply written comments about the quality of the evidence. This section discusses responses to these additional measures and argues that they are broadly consistent with a rational account.

When asked to assess the quality of the available evidence, participants typically find evidence that is consistent with their prior beliefs more convincing than evidence that is inconsistent with their prior beliefs (Lord et al., 1979; Munro & Ditto, 1997; McHoskey, 1995; Taber et al., 2009). For example, Lord et al. (1979) asked participants to rate "how well or poorly" each study had been conducted, and "how convincing" each study seemed "as evidence of the deterrent efficacy of capital punishment" (p. 2101). Responses to these questions differed by around 2.5 points on a 17-point scale depending on whether they were provided by supporters or opponents of the death penalty. Lord et al. (1979) also solicited written comments, and these comments revealed that participants tended to focus more on the methodological flaws of the studies that disputed their prior beliefs. Both the numerical ratings and the written comments therefore suggested that participants let their prior beliefs influence the way they evaluated the evidence, a phenomenon we refer to as *biased evaluation*.

Biased evaluation may be undesirable in some contexts, but there is a general consensus that evaluating evidence in this way can be rational (Gerber & Green, 1999; Koehler, 1993). Lord et al. (1979) appear to have agreed with this consensus and wrote that "there can be no real quarrel with a willingness to infer that studies supporting one's theory-based expectations are more probative than, or methodologically superior to, studies that contradict one's expectations" (p. 2106). They pointed out, for example, that it is reasonable to be skeptical about reports of "miraculous virgin births" and "herbal cures for cancer" and suggested that the biased evaluations reported by their participants can be defended on similar grounds.

## Biased Assimilation of Evidence

Lord et al. (1979) suggested that the main "inferential shortcoming" of their participants "did not lie in their inclination to process evidence in a biased manner." Instead, "their sin lay in their readiness to use evidence already processed in a biased manner to bolster the very theory or belief that initially 'justified' the processing bias" (p. 2107). Lord et al. (1979) referred to this phenomenon as *biased assimilation*, and others have adopted this term to refer to irrational behavior that results in belief divergence (Liberman & Chaiken, 1992; McHoskey, 1995; Munro & Ditto, 1997; Plous, 1991).

Our work challenges the intuition that biased assimilation is always an inferential sin. To see why, consider the model predictions for the forced choice version of the polarization condition in our medical diagnosis experiment. After observing two incompatible test results,

---

[4] These *p*-values for the most common responses were computed using a likelihood ratio test described by Nettleton (2009).

the model carries out a "biased evaluation" and infers that the less likely result is probably spurious. Discounting this spurious result allows the model to strengthen its belief in the diagnosis supported by the other test result. Note, however, that the belief that is strengthened is the very same belief that provided grounds for discounting the spurious result. We therefore believe that the inferences of our normative model are consistent with the notion of biased assimilation.

Although some kinds of biased assimilation are consistent with normative probabilistic inference, there are surely examples of biased assimilation that depart from rational norms. For example, our model discounts the less likely test result in the polarization conditions, but assigns a small probability to the possibility that this test result is accurate. An inferential approach that does not maintain uncertainty in this way is in danger of arriving at conclusions that are overly confident given the available evidence. For example, participants in the death penalty study could have become overconfident if they judged the evidence that was incompatible with their prior beliefs to be spurious and then completely ignored or discarded that evidence, leaving only evidence that supported their prior beliefs.

## Belief Coherence

Biased assimilation tends to ensure that the conclusions people draw are maximally consistent with their beliefs about the specific evidence they observe. Biased assimilation is therefore related to the psychological literature on belief coherence, which proposes that people tend to adjust their beliefs in a way that maximizes the coherence among these beliefs. Some researchers (Simon, Snow, & Read, 2004) have proposed that coherence-based accounts of belief revision are incompatible with normative probabilistic accounts like ours. We now argue, however, that these two views of reasoning can be reconciled.

Simon et al. (2004) suggested that Bayesian models are unable to capture bidirectional reasoning in which "the evidence influences the conclusions and, at the same time, the emerging conclusion affects the evaluation of the evidence" (p. 814). Bidirectional reasoning is often modeled using constraint-satisfaction networks, in which the edges represent ways in which the values of the nodes mutually constrain each other. A classic example is a network in which the nodes represent the corners of the Necker cube, and the edges enforce the notion that the interpretations of these corners should be mutually consistent. Networks of this kind, however, can be viewed as instances of probabilistic models. For example, McClelland (2013, chapter 3) showed how the perception of the Necker cube can be modeled using a Boltzmann machine, which can be viewed as a constraint-satisfaction network that relies on probabilistic inference (Hinton, 2007).

Research using Boltzmann machines and related models demonstrates that there is no fundamental incompatibility between probabilistic reasoning and coherence-based reasoning. Studies of belief coherence, however, have documented many specific phenomena, and case studies are needed to determine whether these individual phenomena are compatible with probabilistic inference. The rest of this section considers one such phenomenon that is closely related to biased assimilation and belief polarization. We focus on coherence shifts, which occur when people's beliefs and attitudes shift to become more coherent with their eventual decisions (Holyoak & Simon, 1999; Simon & Holyoak, 2002; Simon, Pham, Quang, & Holyoak, 2001; Simon et al., 2004). As we see,

some but not all of these shifts appear to be consistent with our probabilistic account.

Holyoak and Simon (1999; Simon et al. 2001) have documented coherence shifts using experiments in which participants act as judges in legal cases. Participants' initial beliefs about the relevant legal issues in each case were relatively neutral. However, upon reviewing the arguments for the case, participants' beliefs polarized to favor the party they eventually ruled in favor of. The experimental paradigm in these studies is very similar to the one employed in many polarization studies: Participants express an initial belief, review some evidence, and then express a final belief. As with previous polarization studies, participants' beliefs in these coherence studies diverged to become more extreme after reviewing the evidence.

As a specific example, consider Experiment 3 of Holyoak and Simon (1999). Participants in this experiment evaluated a case involving a company that was suing an investor who posted a negative message about the company on an online bulletin board. The plaintiff and defendant made arguments that differed with respect to six specific issues, including whether the negative message caused the company to collapse, whether the investor's primary motive was vindictiveness or a desire to protect other potential investors, and whether messages on an online bulletin boards should be treated like newspaper articles or telephone messages for legal purposes. Before reviewing these arguments, participants were given descriptions of the defendant that suggested he was either honest or dishonest. Participants who read the honest description tended to rule in the defendant's favor and to uphold the defendant's position with respect to all six of the disputed issues, and participants who read the dishonest description tended to do the opposite.

Holyoak and Simon (1999) acknowledged that any plausible account of decision making will predict that the descriptions of the defendant's character should affect participants' inferences about his motivation. They suggested, however, that normative accounts will struggle to explain why the character description affects inferences about all six of the disputed issues: "For example, an inference from the shady history of the defendant to the conclusion that the Internet resembles a newspaper more than it does a telephone system is coherent . . . but not logically compelling" (p. 12). We believe, however, that our probabilistic account can explain why the character descriptions affect beliefs about all six disputed issues. The experiment can be modeled using an expanded version of the Bayes net in Figure 3a.vi in which node $H$ is split into six nodes, one for each disputed issue, node $D$ is similarly split into multiple nodes, and node $V$ represents whether the defendant is honest. If the CPDs reflect the fact that dishonest individuals tend to make statements that are not true, learning that the defendant is dishonest will tend to cast doubt on everything that he says, including his claim that bulletin boards are like telephone messages. Similarly, learning that the defendant is honest will tend to lend support to all of his positions with respect to the disputed issues.

Although the coherence shift just described appears to be consistent with our probabilistic account, other examples of coherence shifts are not. In particular, our account is insensitive to the order in which information is presented, and therefore will not capture coherence shifts that rely on order effects (Bond, Carlson, Meloy, Russo, & Tanner, 2007; Russo, Carlson, & Meloy, 2006; Russo, Carlson, Meloy, & Yang, 2008). Order effects of this kind, however, are often consistent with models that rely on approximate

probabilistic inference (Sanborn, Griffiths, & Navarro, 2010). We have argued throughout that belief polarization is consistent with full probabilistic inference, but incorporating algorithms for approximate inference may expand the set of phenomena that our account is able to capture.

## "Hot" Versus "Cold" Cognition

Normative probabilistic inference can be viewed as an idealization of human reasoning. Some inferences, however, are shaped by emotion in addition to reason. Inferences of this kind are sometimes described as examples of "hot" cognition (Abelson, 1963). Previous work on belief polarization has typically focused on hot cognition. For example, the polarization studies described earlier explore issues that are emotionally engaging and of significant personal importance, such as political (Lord et al. 1979; Plous, 1991; Taber e tal. 2009), personal health (Liberman & Chaiken, 1992), and religious (Batson, 1975) beliefs. Theoretical accounts of these experiments typically invoke the notion of motivated reasoning and propose that reasoners make inferences that are distorted in the direction of conclusions that they would like to believe (e.g., Taber et al. 2009; Taber & Lodge, 2006).

In contrast, we have argued that belief polarization can sometimes result from "cold" cognition. Our probabilistic account acknowledges that reasoners may have different beliefs about which hypotheses are likely to be true, but does not require that reasoners differ in their affects toward the hypotheses under consideration. In keeping with this emphasis on cold cognition, our experiment used a scenario that is both abstract and impersonal and therefore unlikely to provoke strong emotion.

## Identifying Irrational Behavior

We have argued that belief polarization, biased assimilation, and coherence shifts may be rational in some contexts. This argument relies on the role of background knowledge in reasoning. Studies of reasoning often aim to reduce or eliminate the influence of background knowledge, but invoking such knowledge comes so naturally to most people that Stanovich (1999) has called this tendency the "fundamental computational bias." Multiple studies have suggested that people's inferences can be explained by taking relevant background knowledge into account. For instance, Oaksford and Chater (1994) demonstrated that people's behavior in the Wason selection task is consistent with a normative account if people assume that most properties are rare. Similarly, people's causal inferences are influenced by their existing knowledge about possible alternative causes and disabling conditions (Cummins et al. 1991). Although one could argue that background knowledge is not relevant to some reasoning phenomena, it is hard to argue that such knowledge is not relevant to belief polarization, particularly because studies of polarization often categorize participants on the basis of their prior beliefs.

This line of reasoning might seem to imply that appropriate background knowledge can always be invoked to classify a given inference as rational under some conditions. It is therefore natural to wonder whether irrational behavior can ever be definitively identified (Cohen, 1981). The primary goal of this article was to show that some apparently questionable inferences may be rational after all, but the same normative approach can be used to identify irrational behavior. For instance, we used our normative account to identify cases in our

experiment in which it would have been irrational to diverge as well as cases in which it would have been irrational *not* to diverge. Although our participants' overall judgments were consistent with our normative account, many individual participants exhibited behavior that was inconsistent with our normative account, such as diverging when this was not the normative response.

More generally, formal probabilistic analyses can be used to characterize other belief revision behaviors that are unequivocally irrational. One example might be called "inevitable belief reinforcement," in which someone updates his or her belief about a hypothesis in the same direction, regardless of the data (e.g., Pitz et al., 1967). A gambler who becomes increasingly convinced that a roulette wheel is biased in favor of red whether the next spin produces red, black, or green, would be showing inevitable belief reinforcement. This behavior is provably inconsistent with any fully normative probabilistic approach, and therefore would provide strong evidence of irrationality.

## Larger Hypothesis Spaces

Throughout this article we assumed that people were reasoning about binary hypothesis spaces (e.g., the death penalty either does or does not deter crime). This assumption is consistent with most psychological studies of belief divergence, which focus on beliefs about two mutually exclusive hypotheses. For some problems, however, like predicting how many Democrats will win in a Congressional election, there are more than two possibilities. This section briefly discusses how polarization can be defined when people are reasoning about large hypothesis spaces.

The most natural extension of the approach developed here is to consider polarization on a hypothesis-by-hypothesis basis. Given any single hypothesis $H$, Equation 1 can be used to characterize whether two people diverge with respect to that hypothesis. Note, however, that it is possible for two people to diverge in their beliefs about one hypothesis (e.g., $H = 1$) while simultaneously converging in their beliefs about another (e.g., $H = 3$).

If $H$ lies on an interval scale, like in the election prediction example above, more global notions of divergence can also be considered. One such definition holds that divergence occurs if the difference between the weighted averages of two people's beliefs about $H$ increases. Although our normative analysis did not consider global belief divergence, political science researchers have developed probabilistic analyses that suggest that global belief divergence can be normative in some circumstances (Dixit & Weibull, 2007; Gerber & Green, 1999).

## Conclusion

This article presented a normative probabilistic approach that can account for belief divergence, a phenomenon that is typically considered to be incompatible with normative accounts. We applied the approach to several classic studies of belief divergence and presented a new experiment that confirms the prediction that manipulating prior beliefs can make divergence more or less likely to emerge. Although we propose that some instances of divergence are compatible with a normative account, we do not suggest that human inferences are always or even mostly rational. Our work suggests, however, that deciding whether a given inference is rational demands careful thought, and often, a formal analysis. In some cases, formal analyses

provide baselines for understanding how people's inferences depart from rational norms. In other cases, formal analyses suggest that apparently irrational inferences make sense once all of the relevant information is taken into account.

# References

Abelson, R. P. (1963). Computer simulation of "hot" cognition. In S. S. Tomkins & S. Messick (Eds.), *Computer simulation of personality: Frontier of psychological theory* (pp. 277–298). New York, NY: Wiley.

Andreoni, J., & Mylovanov, T. (2012). Diverging opinions. *American Economic Journal: Microeconomics, 4,* 209–232.

Austerweil, J. L., & Griffiths, T. L. (2011). Seeking confirmation is rational for deterministic hypotheses. *Cognitive Science, 35,* 499–526. doi:10.1111/j.1551-6709.2010.01161.x

Baron, J. (2008). *Thinking and deciding* (4th ed.). Cambridge, England: Cambridge University Press.

Batson, C. D. (1975). Rational processing or rationalization? The effect of disconfirming information on a stated religious belief. *Journal of Personality and Social Psychology, 32,* 176–184. doi:10.1037/h0076771

Bond, S. D., Carlson, K. A., Meloy, M. G., Russo, J. E., & Tanner, R. J. (2007). Information distortion in the evaluation of a single option. *Organizational Behavior and Human Decision Processes, 102,* 240–254. doi:10.1016/j.obhdp.2006.04.009

Bullock, J. G. (2009). Partisan bias and the Bayesian ideal in the study of public opinion. *Journal of Politics, 71,* 1109–1124. doi:10.1017/S0022381609090914

Camerer, C. F., & Ho, T. H. (1994). Violations of the betweenness axiom and nonlinearity in probability. *Journal of Risk and Uncertainty, 8,* 167–196. doi:10.1007/BF01065371

Cohen, L. J. (1981). Can human irrationality be experimentally demonstrated? *Behavioral and Brain Sciences, 4,* 317–370. doi:10.1017/S0140525X00009092

Cummins, D. D., Lubart, T., Alksnis, O., & Rist, R. (1991). Conditional reasoning and causation. *Memory & Cognition, 19,* 274–282. doi:10.3758/BF03211151

Dawson, E., Gilovich, T., & Regan, D. T. (2002). Motivated reasoning and performance on the Wason selection task. *Personality and Social Psychology Bulletin, 28,* 1379–1387. doi:10.1177/014616702236869

Dixit, A. K., & Weibull, J. W. (2007). Political polarization. *Proceedings of the National Academy of Sciences of the United States of America, 104,* 7351–7356. doi:10.1073/pnas.0702071104

Duhem, P. (1954). *The aim and structure of physical theory*. Princeton, NJ: Princeton University Press.

Edwards, W., & Fasolo, B. (2001). Decision technology. *Annual Review of Psychology, 52,* 581–606. doi:10.1146/annurev.psych.52.1.581

Fryer, R. G., Jr., Harms, P., & Jackson, M. O. (2013, June). *Updating beliefs with ambiguous evidence: Implications for polarization* (Working Paper No. 19114). National Bureau of Economic Research, Cambridge, MA.

Gerber, A., & Green, D. (1999). Misperceptions about perceptual bias. *Annual Review of Political Science, 2,* 189–210. doi:10.1146/annurev.polisci.2.1.189

Gigerenzer, G. (1991). How to make cognitive illusions disappear: Beyond "heuristics and biases". *European Review of Social Psychology, 2,* 83–115. doi:10.1080/14792779143000033

Gilovich, T. D., & Griffin, D. W. (2010). Judgment and decision making. In S. T. Fiske, D. T. Gilbert, & G. Lindzey (Eds.), *Handbook of social psychology* (5th ed., Vol. 1, pp. 542–588). Hoboken, NJ: Wiley.

Gonzalez, R., & Wu, G. (1999). On the shape of the probability weighting function. *Cognitive Psychology, 38,* 129–166. doi:10.1006/cogp.1998.0710

Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., Kushnir, T., & Danks, D. (2004). A theory of causal learning in children: Causal maps and Bayes nets. *Psychological Review, 111,* 3–32. doi:10.1037/0033-295X.111.1.3

Green, C. S., Benson, C., Kersten, D., & Schrater, P. (2010). Alterations in choice behavior by manipulations of world model. *Proceedings of the National Academy of Sciences of the United States of America, 107,* 16401–16406. doi:10.1073/pnas.1001709107

Hahn, U., & Oaksford, M. (2006). A Bayesian approach to informal argument fallacies. *Synthese, 152,* 207–236. doi:10.1007/s11229-005-5233-2

Hahn, U., & Oaksford, M. (2007). The rationality of informal argumentation: A Bayesian approach to reasoning fallacies. *Psychological Review, 114,* 704–732. doi:10.1037/0033-295X.114.3.704

Halpern, J. Y., & Pass, R. (2010). I don't want to think about it now: Decision theory with costly computation. In F. Lin, U. Sattler, & M. Truszczynski (Eds.), *Proceedings of the 12th international conference on principles of knowledge representation and reasoning* (pp. 182–190). Retrieved from http://aaai.org/ocs/index.php/KR/KR2010/paper/view/1353/1616

Harris, A. J. L., Hsu, A. S., & Madsen, J. K. (2012). Because Hitler did it! Quantitative tests of Bayesian argumentation using *ad hominem. Thinking & Reasoning, 18,* 311–343. doi:10.1080/13546783.2012.670753

Heit, E., & Nicholson, S. P. (2010). The opposite of Republican: Polarization and political categorization. *Cognitive Science, 34,* 1503–1516. doi:10.1111/j.1551-6709.2010.01138.x

Hinton, G. E. (2007). Boltzmann machine. *Scholarpedia, 2,* 1668. doi:10.4249/scholarpedia.1668

Holyoak, K. J., & Simon, D. (1999). Bidirectional reasoning in decision making by constraint satisfaction. *Journal of Experimental Psychology: General, 128,* 3–31. doi:10.1037/0096-3445.128.1.3

Jaynes, E. T. (2003). *Probability theory: The logic of science* (G. L. Bretthorst, Ed.). Cambridge, England: Cambridge University Press. doi:10.1017/CBO9780511790423

Jern, A., Chang, K. K., & Kemp, C. (2009). Bayesian belief polarization. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, & A. Culotta (Eds.), *Advances in neural information processing systems* (Vol. 22, pp. 853–861). Red Hook, NY: Curran Associates.

Kahneman, D., Slovic, P., & Tversky, A. (Eds.). (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge, England: Cambridge University Press. doi:10.1017/CBO9780511809477

Kelly, T. (2008). Disagreement, dogmatism, and belief polarization. *The Journal of Philosophy, 105,* 611–633.

Klaczynski, P. A. (2000). Motivated scientific reasoning biases, epistemological beliefs, and theory polarization: A two-process approach to adolescent cognition. *Child Development, 71,* 1347–1366. doi:10.1111/1467-8624.00232

Koehler, J. J. (1993). The influence of prior beliefs on scientific judgments of evidence quality. *Organizational Behavior and Human Decision Processes, 56,* 28–55. doi:10.1006/obhd.1993.1044

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin, 108,* 480–498. doi:10.1037/0033-2909.108.3.480

Liberman, A., & Chaiken, S. (1992). Defensive processing of personally relevant health messages. *Personality and Social Psychology Bulletin, 18,* 669–679. doi:10.1177/0146167292186002

Lopes, L. L., & Ekberg, P. H. (1980). Test of an ordering hypothesis in risky decision making. *Acta Psychologica, 45,* 161–167. doi:10.1016/0001-6918(80)90028-1

Lord, C. G., Ross, L., & Lepper, M. R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology, 37,* 2098–2109. doi:10.1037/0022-3514.37.11.2098

McClelland, J. L. (2013). *Explorations in parallel distributed processing: A handbook of models, programs, and exercises*. Retrieved from http://www.stanford.edu/group/pdplab/pdphandbook/

McHoskey, J. W. (1995). Case closed? On the John F. Kennedy assassination: Biased assimilation of evidence and attitude polarization. *Basic and Applied Social Psychology, 17,* 395–409. doi:10.1207/s15324834basp1703_7

McKenzie, C. R. M. (2004). Framing effects in inference tasks—And why they are normatively defensible. *Memory & Cognition, 32,* 874–885. doi:10.3758/BF03196866

Mullen, B., Dovidio, J. F., Johnson, C., & Copper, C. (1992). In-group–out-group differences in social projection. *Journal of Experimental Social Psychology, 28,* 422–440. doi:10.1016/0022-1031(92)90040-Q

Munro, G. D., & Ditto, P. H. (1997). Biased assimilation, attitude polarization, and affect in reactions to stereotype-relevant scientific information. *Personality and Social Psychology Bulletin, 23,* 636–653. doi:10.1177/0146167297236007

Navarro, D. J., & Perfors, A. F. (2011). Hypothesis generation, sparse categories, and the positive test strategy. *Psychological Review, 118,* 120–134. doi:10.1037/a0021110

Nettleton, D. (2009). Testing for the supremacy of a multinomial cell probability. *Journal of the American Statistical Association, 104,* 1052–1059. doi:10.1198/jasa.2009.tm08213

Nisbett, R. E., & Ross, L. D. (1980). *Human inference: Strategies and shortcomings of social judgment.* Englewood Cliffs, NJ: Prentice-Hall.

Nisbett, R. E., Zukier, H., & Lemley, R. E. (1981). The dilution effect: Nondiagnostic information weakens the implications of diagnostic information. *Cognitive Psychology, 13,* 248–277. doi:10.1016/0010-0285(81)90010-4

Oaksford, M., & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review, 101,* 608–631. doi:10.1037/0033-295X.101.4.608

Oaksford, M., & Hahn, U. (2004). A Bayesian approach to the argument from ignorance. *Canadian Journal of Experimental Psychology, 58,* 75–85. doi:10.1037/h0085798

O'Connor, B. (2006). *Biased evidence assimilation under bounded Bayesian rationality* (Unpublished master's thesis). Stanford University, Stanford, CA.

Pearl, J. (2000). *Causality: Models, reasoning, and inference.* Cambridge, England: Cambridge University Press.

Phillips, L. D., & Edwards, W. (1966). Conservatism in a simple probability inference task. *Journal of Experimental Psychology, 72,* 346–354. doi:10.1037/h0023653

Pitz, G. F., Downing, L., & Reinhold, H. (1967). Sequential effects in the revision of subjective probabilities. *Canadian Journal of Psychology, 21,* 381–393. doi:10.1037/h0082998

Plous, S. (1991). Biases in the assimilation of technological breakdowns: Do accidents make us safer? *Journal of Applied Social Psychology, 21,* 1058–1082. doi:10.1111/j.1559-1816.1991.tb00459.x

Pomerantz, E. M., Chaiken, S., & Tordesillas, R. S. (1995). Attitude strength and resistance processes. *Journal of Personality and Social Psychology, 69,* 408–419. doi:10.1037/0022-3514.69.3.408

Putnam, H. (1974). The "corroboration" of theories. In P. A. Schilpp (Ed.), *The philosophy of Karl Popper* (pp. 121–137). La Salle, IL: The Open Court.

Quine, W. V. O. (1951). Two dogmas of empiricism. *Philosophical Review, 60,* 20–43.

Rabin, M., & Schrag, J. L. (1999). First impressions matter: A model of confirmatory bias. *Quarterly Journal of Economics, 114,* 37–82. doi:10.1162/003355399555945

Rehder, B., & Hastie, R. (1996). The moderating influence of variability on belief revision. *Psychonomic Bulletin & Review, 3,* 499–503. doi:10.3758/BF03214554

Ross, L., & Anderson, C. A. (1982). Shortcomings in the attribution process: On the origins and maintenance of erroneous social assessments. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 129–152). Cambridge, England: Cambridge University Press. doi:10.1017/CBO9780511809477.010

Ross, L., & Lepper, M. R. (1980). The perseverance of beliefs: Empirical and normative considerations. In R. A. Shweder (Ed.), *New directions for methodology of social and behavioral science: Fallible judgment in behavioral research* (pp. 17–36). San Francisco, CA: Jossey-Bass.

Russo, J. E., Carlson, K. A., & Meloy, M. G. (2006). Choosing an inferior alternative. *Psychological Science, 17,* 899–904. doi:10.1111/j.1467-9280.2006.01800.x

Russo, J. E., Carlson, K. A., Meloy, M. G., & Yang, K. (2008). The goal of consistency as a cause of information distortion. *Journal of Experimental Psychology: General, 137,* 456–470. doi:10.1037/a0012786

Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2010). Rational approximations to rational models: Alternative algorithms for category learning. *Psychological Review, 117,* 1144–1167. doi:10.1037/a0020511

Sher, S., & McKenzie, C. R. M. (2008). Framing effects and rationality. In N. Chater & M. Oaksford (Eds.), *The probabilistic mind: Prospects for Bayesian cognitive science* (pp. 79–96). Oxford, England: Oxford University Press. doi:10.1093/acprof:oso/9780199216093.003.0004

Simon, D., & Holyoak, K. J. (2002). Structural dynamics of cognition: From consistency theories to constraint satisfaction. *Personality and Social Psychology Review, 6,* 283–294. doi:10.1207/S15327957PSPR0604_03

Simon, D., Pham, L. B., Quang, A. L., & Holyoak, K. J. (2001). The emergence of coherence over the course of decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27,* 1250–1260. doi:10.1037/0278-7393.27.5.1250

Simon, D., Snow, C. J., & Read, S. J. (2004). The redux of cognitive consistency theories: Evidence judgments by constraint satisfaction. *Journal of Personality and Social Psychology, 86,* 814–837. doi:10.1037/0022-3514.86.6.814

Sloman, S. A. (2005). *Causal models: How people think about the world and its alternatives.* New York, NY: Oxford University Press. doi:10.1093/acprof:oso/9780195183115.001.0001

Stanovich, K. E. (1999). *Who is rational? Studies of individual differences in reasoning.* Mahwah, NJ: Erlbaum.

Taber, C. S., Cann, D., & Kucsova, S. (2009). The motivated processing of political arguments. *Political Behavior, 31,* 137–155. doi:10.1007/s11109-008-9075-8

Taber, C. S., & Lodge, M. (2006). Motivated skepticism in the evaluation of political beliefs. *American Journal of Political Science, 50,* 755–769. doi:10.1111/j.1540-5907.2006.00214.x

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representations of uncertainty. *Journal of Risk and Uncertainty, 5,* 297–323. doi:10.1007/BF00122574

Wu, G., & Gonzalez, R. (1996). Curvature of the probability weighting function. *Management Science, 42,* 1676–1690. doi:10.1287/mnsc.42.12.1676

Zimper, A., & Ludwig, A. (2009). On attitude polarization under Bayesian learning with non-additive beliefs. *Journal of Risk and Uncertainty, 39,* 181–212. doi:10.1007/s11166-009-9074-0

## Appendix A

## Proof That Family 1 Bayes Nets Cannot Produce Contrary Updating

In the main text, we claimed that the Bayes nets in Figure 3a can be organized into two families. This section explains in detail why the networks in Family 1 cannot produce contrary updating. The main text described how the value of the likelihood ratio $\frac{P_A(D \mid H = 1)}{P_A(D \mid H = 2)}$ determines the direction in which A's beliefs will change. Equation 1 specified a general criterion for contrary updating; our discussion of likelihood ratios in the main text implies that when there are only two possible hypotheses, this criterion is equivalent to the following:

$$\frac{P_A(D \mid H = 1)}{P_A(D \mid H = 2)} > 1 \quad \text{and} \quad \frac{P_B(D \mid H = 1)}{P_B(D \mid H = 2)} < 1, \text{ or vice versa.}$$
(A1)

So far, we have maintained two distributions, $P_A(\cdot)$ and $P_B(\cdot)$, for the two people. Our analysis can be simplified by using a single distribution $P(\cdot)$ for both people and adding a background knowledge node $\beta$ that captures the differences between A and B. Because we assume that the two people differ only with respect to their prior beliefs about the root nodes, $\beta$ can be viewed as a switch that sets the beliefs for these nodes. The result of adding the $\beta$ node to our Bayes nets is shown in Figure A1a. By conditioning on $\beta$, the likelihood ratio for Person A can be rewritten as

$$\frac{P_A(D \mid H = 1)}{P_A(D \mid H = 2)} = \frac{P(D \mid H = 1, \beta = A)}{P(D \mid H = 2, \beta = A)}.$$
(A2)

Figure A1b shows how an expanded version of the Bayes net in Figure 3b can accommodate the beliefs of two different people. In the expanded Bayes net, the value of $\beta$ determines the prior beliefs about $H$. When $\beta = A$, this Bayes net behaves identically to the one in Figure 3b. However, the expanded version of the Bayes net can also capture the different prior beliefs of Person B: when $\beta = B$, more prior probability is assigned to $H = 2$ than to $H = 1$.

Equation A1 implies that contrary updating cannot occur if the likelihood ratios for two people are identical. We can use this observation to establish that contrary updating is impossible if either of the following two conditions holds:

C1: $D$ and $\beta$ are conditionally independent given $H$.

C2: $D$ and $H$ are conditionally independent given $\beta$.

First consider condition C1, which captures the idea that $\beta$ provides no information about $D$ once $H$ is known. When this condition is met, the likelihood ratio on the right of Equation A2
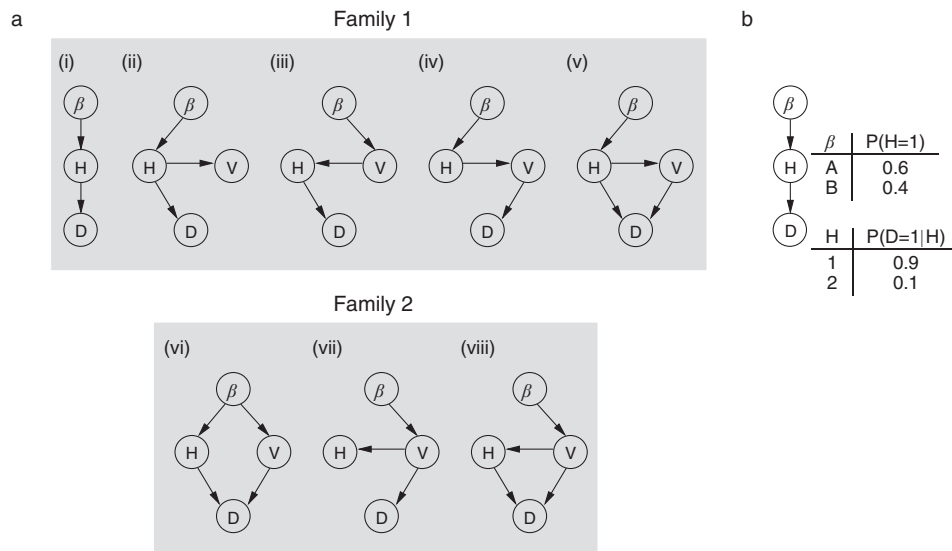


*Figure A1.* Expanded versions of the Bayes nets in Figure 3. Node $\beta$ acts as a switch that assigns different prior probabilities to the root nodes of the Bayes nets in Figure 3 for different people. H = hypothesis; D = data; V = variable.

(*Appendices continue*)

reduces to $\dfrac{P(D \mid H = 1)}{P(D \mid H = 2)}$. Because this reduced term does not depend on $\beta$, it must be the same for both people, which means that contrary updating cannot occur. Condition C1 applies to all of the Bayes nets in Family 1. For each network in this family, note that there are no paths from $\beta$ to $D$ that do not pass through $H$. As a result, $D$ is independent of $\beta$ if the value of $H$ is known. This independence relationship means that if $H = 1$, then the two people have identical expectations about the value of $D$, and similarly if $H = 2$. As a result, the two people must make identical inferences about how data $D$ bear on hypothesis $H$. Condition C1 does not apply to the Bayes nets in Family 2, and each of these networks allows background knowledge to influence how the data $D$ are interpreted. As a result, both contrary and parallel updating are possible for Bayes nets in Family 2.

Now consider condition C2, which captures the idea that $D$ provides no information about $H$ once $\beta$ is known. If this condition

holds, then the likelihood ratio on the right of Equation A2 reduces to $\dfrac{P(D \mid \beta = A)}{P(D \mid \beta = A)} = 1$, and observing $D$ does not lead either person to update his or her beliefs about $H$. Because we have focused on cases in which the variables $H$, $D$, and $V$ are all linked in some way, condition C2 does not apply to any of the Bayes nets in Figure A1a.

Our analysis has focused on three-node Bayes nets that include variable $V$ in addition to $D$ and $H$. For some purposes it may be necessary to consider networks with four or more nodes that include multiple variables $V_i$ in addition to $D$ and $H$. Conditions C1 and C2 both apply to Bayes nets of any size, and the arguments above imply that contrary updating is impossible if either condition holds. We conjecture that any network not excluded by these conditions can produce contrary updating for some settings of its CPDs.

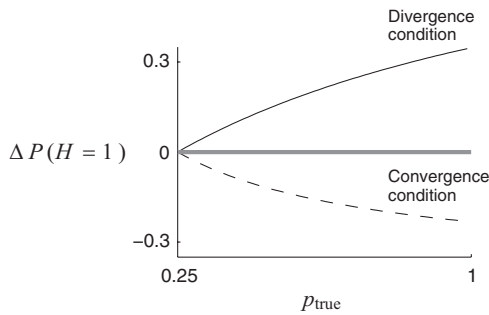# Appendix B

## Model Prediction Details



*Figure B1.* Results of the first sensitivity analysis described in the text. The plot shows belief change as a function of the $p_{\text{true}}$ parameter. Although the magnitudes of belief change increase with $p_{\text{true}}$, the direction of change remains the same for all valid values of the parameter.

The main text states that model predictions for the diagnosis scenario in our experiment do not depend on the exact parameters used in the CPD for $P(D \mid V)$. This section describes two analyses that support this conclusion.

In both analyses, each test is most likely to indicate the true disease. In one analysis, we assumed that the tests would falsely indicate the remaining three diseases with equal probability. That is, we assumed that each test would indicate the true disease with probability $p_{\text{true}}$ and that the three remaining diseases would each be indicated with probability $p_{\text{false}}$, with $p_{\text{true}} > p_{\text{false}}$. We computed model predictions for all sets of probabilities meeting the specified constraints, in increments of .01. The results of this analysis are shown in Figure B1. Although the magnitude of change varied, the predicted change in belief was always in the direction reported in the text.

In the second analysis, we assumed that the tests were more likely to indicate diseases in the same class as the true disease. That is, we assumed that each test would indicate the wrong disease of the correct class with probability $p_{\text{false}_1}$ and would indicate each of the remaining two diseases with probability $p_{\text{false}_2}$, with $p_{\text{true}} > p_{\text{false}_1} > p_{\text{false}_2}$. Once again, we computed model predictions for all sets of probabilities meeting the specified constraints, in increments of .01, and the predicted change in belief was always in the reported direction.

The main text also states that the model predictions hold if the prior distributions are distorted according to a weighting function like that used by prospect theory. We verified this claim using two weighting functions. The first is a one-parameter weighting function supported by several studies of choice behavior (Camerer & Ho, 1994; Tversky & Kahneman, 1992; Wu & Gonzalez, 1996):

$$w(p) = \frac{p^{\beta}}{(p^{\beta} + (1 - p)^{\beta})^{1/\beta}},$$

where $p$ is the true probability and $w(p)$ is the distorted probability. We verified the model predictions using $\beta = 0.56$ (Camerer & Ho, 1994), $\beta = 0.61$ (Tversky & Kahneman, 1992), and $\beta = 0.71$ (Wu & Gonzalez, 1996). The second weighting function was a two-parameter function proposed by Gonzalez and Wu (1999):

$$w(p) = \exp(-\delta(-\log(p))^{\gamma}),$$

with $\delta = 0.77$ and $\gamma = 0.44$.