Biased evidence assimilation under bounded Bayesian rationality

Brendan O'Connor

Revised September 5, 2006

Master Thesis, Symbolic Systems Program, Stanford University

For an individually designed track in Computation, Cognition, and Social Science

I certify that I have read this thesis and it is adequate as partial fulfillment of the requirements for the Master of Science in Symbolic Systems.

(Signed via email, 8/30/2006)

Jonathan Bendor (advisor)

James McClelland (second reader)

Contents

1 Abstract

I explain evidence assimilation bias as the result of agents trying to maintain cognitive consistency. This can be interpreted as a boundedly rational inference method — local search for a maximally likely world model. Given a sufficiently complex network of beliefs, such an approximate Bayesian can display systematically non-Bayesian behavior. These arguments are first sketched via connectionist Hopfield networks, in line with previous psychology literature, and then illustrated and analyzed in more detail with probabilistic graphical models — Bayesian networks and Markov random fields.

2 Introduction

Any theory of how people update their beliefs based on evidence needs to account for at least three behaviors. The two basic ones are:

- Belief persistence: People can persist in their previously held beliefs despite new evidence to the contrary.
- Belief revision: People can change their beliefs with sufficiently strong new evidence.

Bayesian probability theory prescribes an optimal balance of persistence and revision. A Bayesian updates his beliefs to become more consistent with the likelihood of the evidence (revision), but moderates this update according to his prior beliefs (persistence).

In many cases, people face complex, ambiguous evidence that must be *interpreted* to determine its compatibility with one's beliefs. This case gives rise to the third more interesting behavior.

- Prior-dependent interpretation: How people interpret evidence depends on their other beliefs.

This implies that different agents can interpret the same piece of evidence in different ways. One particularly pathological example is when agents interpret evidence as supporting their previously held beliefs, which can even cause them to update in different directions; this can result in a polarization of belief among heterogenous group. In general, when agents tend to interpret evidence in a manner favorable to their prior beliefs, the effect may be called *evidence assimilation bias* (Lord, Ross, and Lepper, 1979).

I focus on the problem of evidence with multiple possible interpretations, so the weight of the evidence itself requires a judgment. The correct Bayesian inference on this interpretation variable actually should depend on one's prior beliefs on the hypothesis under question. In fact, the interpretation inference depends on the interpretations of all other pieces of evidence.

A Bayesian needs to maintain a belief distribution over all possible *combinations* of values of variables, not just a single belief level per variable. I argue the computational costs involved are too high to be realistic. A simple, plausible alternative, trying to maintain the most likely explanation and incrementally revising in the light of new evidence, displays order dependence and assimilation bias.

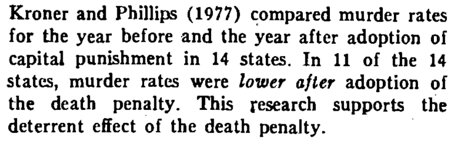
Briefly, the model is as follows: an agent, having initially seen positive evidence, will be rationally skeptical of later negative evidence. But if a large amount of negative evidence accumulates, it would be better to believe in the negative evidence and deny the positive evidence. However, an agent lacking the memory, attention, or motivational resources to review and revise past interpretations will instead continue to make local revisions, becoming stuck in a suboptimal local maximum.

In Section 4 I review some of the connectionist constraint satisfaction models of belief revision. These Hopfield networks display interesting properties, including all three behavioral desiderata listed above. Section 5 reinterprets these models as probabilistic inference, so their behavior can be compared to optimal Bayesian updating. Such a probabilistic graphical model for the evidence interpretation problem is developed and analyzed.

3 Empirical demonstrations of belief persistence bias

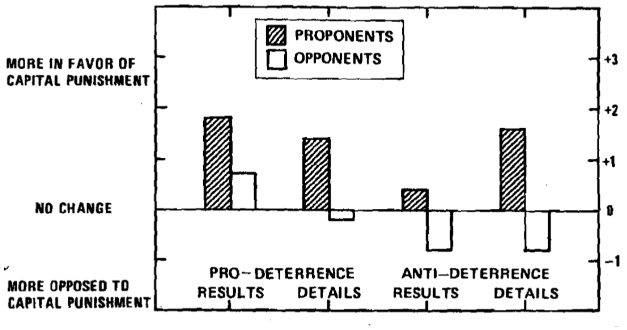
Lord et al. (1979) presented death penalty proponents and opponents with the results and details of supposed empirical studies on whether capital punishment caused a change in crime rates. Like most actual studies on the topic, the evidence was often mixed and open to interpretation. Death penalty proponents tend to find methodological faults in anti-deterrence evidence, and assess pro-deterrence evidence positively; and vice-versa for opponents. In fact, this difference in scrutiny carries over to the final revision on their beliefs — being presented with the same evidence caused participants' self-reported beliefs to *diverge*, as seen in Figure 1, though doubt has been raised on the robustness of the polarization phenomenon (Miller et al., 1993).

Figure 1: (a) Example of evidence used by Lord et al. (1979). (b) Participants' self-reported belief revision after seeing the pro-deterrence then anti-deterrence evidence, and seeing the two-sentence summary versus several pages of details about the study. More information causes more polarization.



Kroner and Phillips (1977) compared murder rates for the year before and the year after adoption of capital punishment in 14 states. In 11 of the 14 states, murder rates were *lower after* adoption of the death penalty. This research supports the deterrent effect of the death penalty.

(a)



(b)

This result seems troubling because it suggests the difficulty of agreement on complex social and political issues. It also seems counter-normative to give evidence different levels of scrutiny. In this paper I will focus on the problem of bias when interpreting evidence relating to a central hypothesis. But numerous other examples of belief persistence biases and other related phenomena exist in the literature, such as confirmation bias,[1] belief perseverance, myside bias, selection bias, cognitive dissonance, order effects, illusory correlations, Einstellung effects, etc. Organizing and summarizing these effects is a daunting task in itself. For that, I defer to reviews such as the ones contained in Nickerson (1998), Rabin and Schrag (1999), Baron ch. 9 (2000), and Nisbett and Ross (1980).

4 Connectionist cognitive consistency

In this section I will first explain the idea of cognitive consistency, then sketch a general Hopfield network model intended be representative of a series of models mostly in the social psychology literature, which I review. This model is interesting for the behavior it exhibits, but I do not analyze it extensively, nor do I specify its psychological interpretation very well; my major model appears in Section 5, using Bayesian probability as its semantics.

4.1 Introduction

A way to conceptualize the structure and relations among a person's beliefs is as a network. Nodes represent individual beliefs or attitudes. Links between nodes can represent a variety of relationships, including relatedness, type inheritance, or other structural aspects of knowledge.

The literature on cognitive consistency treats connections between nodes as indicators of compatibility or degree of association. If two beliefs are strongly connected, such between as BUSH and CONSERVATIVE, there is a positive (consonant) association between them. A negative (dissonant)

[1] The term "confirmation bias", especially historically within the psychology literature, refers to a different, if possibly related, phenomenon: a search strategy biased towards finding evidence confirming one's prior beliefs (Wason, 1977, 1968). But more recently, the term seems to have become more general, encompassing evidence processing biases as well. This use can be seen in psychology (Nickerson, 1998), behavioral economics (Rabin and Schrag, 1999), and numerous popular sources, such as the Wikipedia entry (`http://en.wikipedia.org/wiki/Confirmation_bias`).

connection, such as between BUSH and GAY RIGHTS, indicates a mutual incompatibility.

A possible explanation for belief persistence bias is the theory of cognitive consistency: people strive to maintain consistency among their beliefs (Abelson et al., 1968; Heider, 1958). A person may hold a positive or negative attitude towards each belief node. Cognitive dissonance occurs when a person has two positive, or two negative attitudes towards two negatively linked beliefs; cognitive consonance (consistency) occurs in the opposite case, when a person hold positive attitudes towards consonant beliefs.

Furthermore, cognitive dissonance theory (Festinger, 1957) holds that when people have dissonance, they are motivated to decrease it. This could entail searching for consonant evidence, or changing one's beliefs to fall in line with one another.[2]

Despite the fact that the exact meaning of the positive and negative links can be somewhat vague, psychologists have vividly formalized cognitive consistency theories as Hopfield belief networks, in which consistency is achieved through a soft constraint satisfaction process. Putting specific semantics of these networks on hold, what follows is a description of a simple mathematical model intended to be representative of constraint satisfaction models.

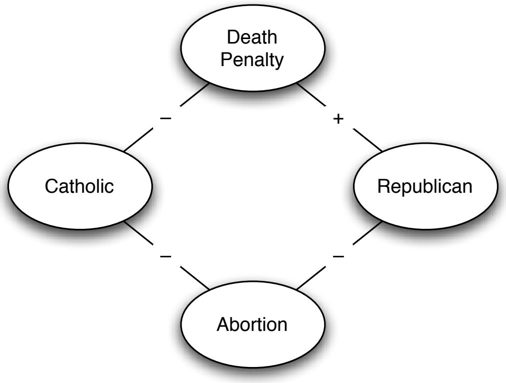
For each belief, an agent holds an evaluation $x_i \in [-1, 1]$ representing the range of dislike/rejection to liking/acceptance. (Perhaps x_i represents affective evaluation, or degree of belief.) A Hopfield network connects these beliefs via bidirectional symmetric weights $w_{ij} = w_{ji} \in \mathbb{R}$. Weights are fixed. We measure the consonance (negative "energy") of a network state as $-E(\vec{x}) = \sum_{i<j} x_i x_j w_{ij}$.[3] Positive links contribute consonance when their endpoints are the same state, and negative links contribute consonance when their endpoints are opposite.

Example Consider Figure 2, specifying the signs of binary constraints over the nodes (C,D,R,A). Any assignment of non-zero values to the nodes violates at least one of the constraints. For example, $\vec{x} = (C, D, R, A) =$

[2] The psychological literature using the term "cognitive dissonance" tends to focus on dissonance between people's self-perceptions and decisions, especially through experiments that force participants to take certain (often bizarre) actions, e.g. Elliot and Levine (1994), Coooper and Fazio (1984); see Lepper and Shultz (2001) for a review, and Zimbardo et al. (1965) to read about poor psych undergrads manipulated into eating grasshoppers. This paper avoids the case of whether decisions conform with beliefs, and concentrates just on the relationships between different beliefs.

[3] The negative sign on $-E$ is annoying but conventional; it is due to the model's derivation from statistical mechanics.

Figure 2: A hard-to-satisfy consistency network. Exact weight values are omitted; which assignment on (C,D,R,A) is optimal depends on the relative magnitude of the weights.



$(1, 1, 1, -1)$ (i.e., conservative Catholic) violates the negative constraint w_{CD}, but fulfills all the other constraints. The consonance of an assignment $\vec{x}$ is

$$-E(\vec{x}) = x_C x_D w_{CD} + x_D x_R w_{DR} + x_R x_A w_{RA} + x_A x_C w_{AC}$$

Table 4.1 illustrates the consonance of several different $\vec{x}$ assignments, assuming all weights are either $+10$ or -10; that is, $(w_{CD}, w_{DR}, w_{RA}, w_{AC}) = (-10, +10, -10, -10)$.

Table 1: Various $\vec{x}$ assignments and their consonance calculations, assuming weight magnitudes of 10. Note the second row cannot be arrived at via linear threshold updates, since it has a 0 value.

x_C	$x_C x_D w_{CD}$	x_D	$x_D x_R w_{DR}$	x_R	$x_R x_A w_{RA}$	x_A	$x_A x_C w_{AC}$	$-E(\vec{x})$
1	-10 (vio)	1	$+10$ (sat)	1	$+10$ (sat)	-1	$+10$ (sat)	20
1	0 (neutral)	0	0 (neutral)	1	$+10$ (sat)	-1	$+10$ (sat)	20
1	-10 (vio)	1	-10 (vio)	-1	$+10$ (sat)	1	-10 (vio)	-20

The constraint satisfaction problem is to find a belief state $\vec{x}$ that maximizes consonance. Each node has a net input activation denoted $a_i =$

$\sum_j w_{ij} x_j$. Writing $-E(\vec{x}) = \frac{1}{2} \sum_i x_i a_i$ makes it apparent that the local consonance contribution of an individual neuron i is proportional to its state multiplied by its net input activation. Hopfield's original formulation (1982) iterates through all nodes, updating them via a linear threshold rule that greedily optimizes consonance:

$$x_i := \begin{cases} -1 & \text{if } a_i < 0 \\ +1 & \text{if } a_i > 0 \\ x_i & \text{if } a_i = 0 \end{cases}$$

So when updating node x_i, the algorithms selects the state for x_i that maximizes consonance. Because the list of neuron states $\vec{x}$ can be thought of as a vector in an n-dimensional space, this algorithm is also called *coordinate ascent* on the consonance function. Since $-E$ is bounded above and weakly increases at each update, the algorithm must converge to a stable state $\vec{x}$ that locally maximizes $-E(\vec{x})$.

It is helpful to think of a Hopfield network as solving a set of soft constraints. For every pair of connected nodes (i, j), there exists a binary constraint function $\phi_{ij}(x_i, x_j) = w_{ij} x_i x_j$. The function ϕ_{ij} can be thought of as a negative cost, "happiness," or consonance, for the current state of the pair (x_i, x_j). The total consonance to be optimized is the sum of all constraint functions. A Hopfield network only uses binary symmetric constraints, but it is easy to conceive of more general constraint satisfaction problems involving non-symmetric or non-binary ϕ functions. (Section 5 develops a probabilistic evidence interpretation task as a problem with many three-way constraints.)

In contrast to the discrete linear threshold rule, a continuous Hopfield network uses a soft S-shaped update function, so the x_i's achieve intermediate values between -1 and $+1$, representing levels of attitude or belief. Artificial neural networks often use the logistic sigmoid $g(a) = 1/(1 + e^{-a})$; for our formulation on $[-1, 1]$ it is useful to use the rescaled tanh form

$$x_i := \tanh(\beta a_i) \quad (= g(2\beta a_i) \times 2 - 1)$$

where the parameter β controls the threshold sharpness; $\beta \to \infty$ approaches the linear threshold rule. There is no guarantee that every update increases consonance, but convergence does occur under asynchronous updating (see MacKay ch. 42 (2003) and Section 5.8). While this rule is less locally greedy than the linear threshold rule, it can still get stuck in local maxima.

Hopfield networks have been used to simulate a number of experiments within the cognitive dissonance tradition, focusing on self-perception and af-

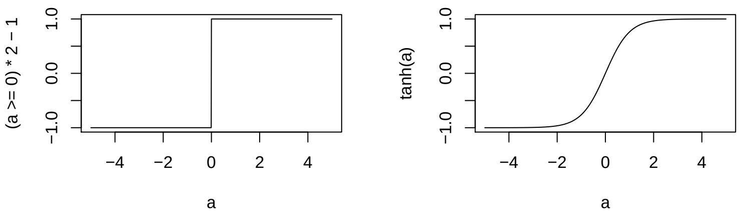


Figure 3: Linear threshold and tanh update functions, $\beta = 1$

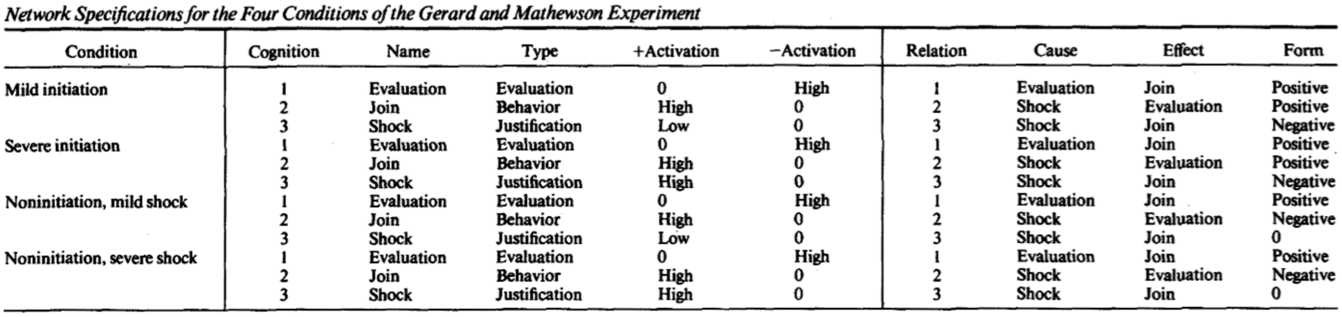
fect (e.g. Shultz and Lepper, 1996; Read et al., 1997). Consistency maximization has been used to analyze other phenomena, such as schemas (Rumelhart et al., 1986; Smolensky, 1986). There are a few analyses of evidence processing.

To give a flavor of some of these models, Figure 4 shows several network configurations used by Shultz and Lepper (1996) to simulate a classic dissonance experiment, in which it was found that when joining a group, participants engaging in embarrassing initiation tasks ended up liking the group more. Shultz and Lepper set up three-node networks whose weights and initial states were intended to represent the different conditions in the experiment. They ran constraint satisfaction updates[4] over many iterations until the node states stabilized, representing participants' dissonance reduction processes.

There has been some work analyzing belief revision from the viewpoint of connectionist constraint satisfaction. Simon et al. (2004) found a tendency toward consistent judgments among multiple pieces of evidence in mock jury experiments, though they did no explicit modeling. Lodge and Taber (2000) describe an affective association explanation for an implicit attitudes study of political beliefs. Paul Thagard's ECHO system — which basically constructs and runs Hopfield constraint satisfaction networks under a variety of possible rules (Thagard, 2002) — has been used to simulate several belief revision examples, including the revision of naive physics beliefs when learning contradictory ideas in elementary physics (Ranney and Thagard, 1988), and “bidirectional reasoning” between arguments and de-

[4] The update rule was different and more complicated than rules presented here. A key feature was that node states would update only incrementally from their current state. “Resistance” parameters similar to β were used, so convergence took many iterations.

Figure 4: One network setup from Shultz and Lepper (1996), simulating the dissonance experiment of Gerard and Mathewson (1966).



Network Specifications for the Four Conditions of the Gerard and Mathewson Experiment

Condition	Cognition	Name	Type	+Activation	−Activation	Relation	Cause	Effect	Form
Mild initiation	1	Evaluation	Evaluation	0	High	1	Evaluation	Join	Positive
	2	Join	Behavior	High	0	2	Shock	Evaluation	Positive
	3	Shock	Justification	Low	0	3	Shock	Join	Negative
Severe initiation	1	Evaluation	Evaluation	0	High	1	Evaluation	Join	Positive
	2	Join	Behavior	High	0	2	Shock	Evaluation	Positive
	3	Shock	Justification	High	0	3	Shock	Join	Negative
Noninitiation, mild shock	1	Evaluation	Evaluation	0	High	1	Evaluation	Join	Positive
	2	Join	Behavior	High	0	2	Shock	Evaluation	Negative
	3	Shock	Justification	Low	0	3	Shock	Join	0
Noninitiation, severe shock	1	Evaluation	Evaluation	0	High	1	Evaluation	Join	Positive
	2	Join	Behavior	High	0	2	Shock	Evaluation	Negative
	3	Shock	Justification	High	0	3	Shock	Join	0

cisions (Holyoak and Simon, 1999). Very little of this work has contrasted constraint network behavior versus Bayesian belief revision,[5] and in general has refrained from analysis of the properties of constraint satisfaction systems, though Read et al. (1997) point out interesting parallels between connectionist system dynamics and the Gestalt psychological theory from which cognitive consistency theory was originally derived.

To my knowledge, have been no attempts to model the Lord et al. (1979) death penalty study. Figure 5 provides a simple illustration in which beliefs are constrained to be consistent with one another. The initial network shows two prior beliefs that the capital punishment deterrent works — these might include remembered anecdotes, analogies from personal experiences, etc. Since they are consistent with one another, there is a positive link between them. This agent believes in both of them. Under the linear threshold rule, this belief state is stable. Next, the agent learns of a new anti-deterrence study. Its negative links with the pro-deterrence studies indicates they contradict, or somehow go against one another. If the agent starts with an initially positive assessment of the study, this introduces a dissonance to the system.

To restore consonance, either the assessment of pro-deterrence evidence has to flip negative, or the assessment of the anti-deterrence study has to flip positive. In this case, updates to either of the pro-deterrence nodes

[5] For example, Simon et al. are excessively hostile to the "algebraic" Bayesian approach, claiming it cannot model complex relations between evidence judgments and hypotheses. This is true of very simple probabilistic models, but as demonstrated in Section 5, graphical model techniques can easily represent, and be used to analyze, such situations. Thagard (2000, 2002) notes a few of the similarities between his constraint satisfaction system and certain types of probabilistic graphical networks.

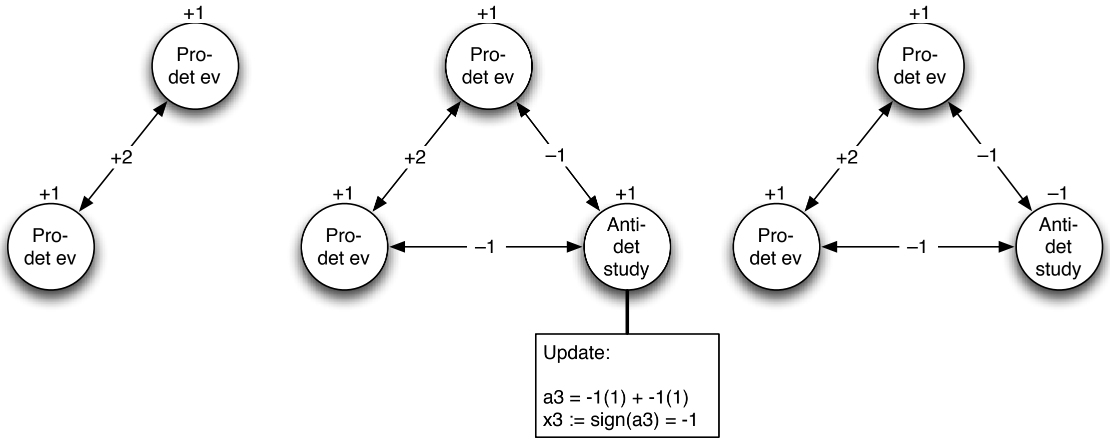


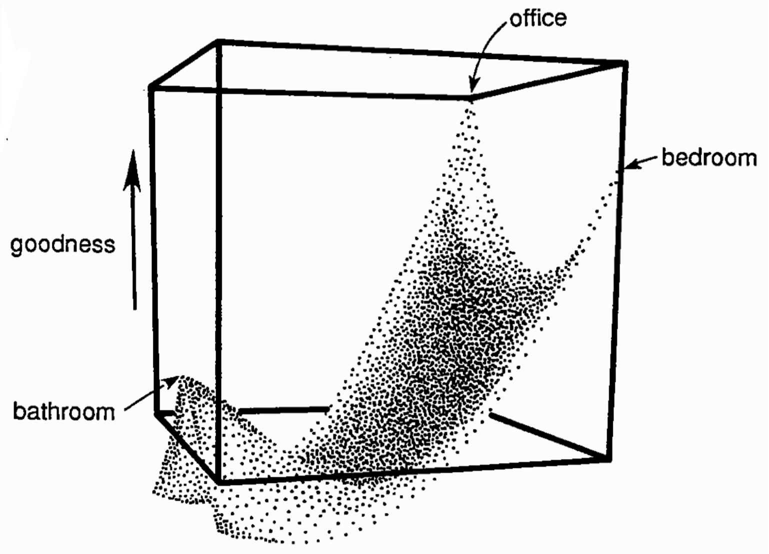
Figure 5: Stable configuration, then weakly dissonant new evidence. A linear threshold update of the new evidence restores stability.

will fail to change them, but updating the anti-deterrence node will flip it negative from the negative influence from the combined weight of multiple pro-deterrence beliefs. A negative belief in the anti-deterrence study is stable.

An interesting analysis of a large, data-derived constraint satisfaction network appears in an early connectionist experiment of Rumelhart et al. (1986), which models schematic concepts as high consonance combinations of features. They first collected judgments from sixteen people of whether 40 room descriptors such as *television*, *bed*, or *windows* were typical of five different room types. The 40 descriptors were connected in a Hopfield network, with the weights w_{ij} high if the descriptors i and j tended to be both on or both off, and w_{ij} low if they tended to be opposite values. The original five room types were evident as local maxima on the consonance surface over the 40-dimensional state space. Given a partial description of a room, constraint satisfaction updates could be run on the missing values to settle on a stable local maximum of what a typical room satisfying that description could be.[6]

[6] In the probabilistic terminology of Section 5, this network was executing hill-climbing MAP estimation conditional on the values contained in the partial description. A number of probabilistic graphical models for representing schema and relations have since been developed, e.g. probabilistic frame-based systems (Koller and Pfeffer, 1998), and Markov logic networks (Domingos et al., 2006). I suspect, but have not worked out in detail, that the room descriptor Hopfield net can be represented as an MLN. Perhaps anticipating

Figure 6: Two-dimensional slice of the room descriptor consonance surface, taken from Rumelhart et al. (1986).



Similarities and differences among the typical room prototypes are evident in the consonance surface; for example, room descriptor configurations corresponding to *office* and *bedroom* are fairly similar, but *bathroom* is very different; combinations of the three are very dissonant, as evidenced by the dip in Figure 6.

We could imagine a belief revision problem occurring in this framework — a person trying categorizing an object based on the incremental processing of evidence. Asch (1946) found that presenting different orders of personality traits can give distinctly different impressions of a person, e.g. "intelligent, industrious, impulsive, critical, stubborn, envious" versus "envious, stubborn, critical, impulsive, industrious, intelligent." The first list caused more favorable impressions than the second, though the only difference is order. Asch's interpretation was that the term "intelligent" colored the interpretation of later evidence. Receiving evidence and forming impressions can be thought of as a search on a consonance surface similar to Rumelhart et al.'s: starting off with "intelligent" as opposed to "envious" initializes your constraint satisfaction search to a different region of possible local maxima; the amount of movement the next several adjectives can cause

these developments, Rumelhart et al. use a Bayesian-justified rule to fit the weights to data.

is limited by your starting position.

4.2 A few properties of Hopfield cognitive consistency networks

In this section I will note a few behavioral properties of cognitive consistency networks that make them attractive models of human reasoning, with the caveat that I'm still too vague about the meaning of node activations and link weights. Section 5 presents a model with much more clearly defined semantics that has many of the same properties.

First, a Hopfield network allows both *persistence* and *revision*. We have already shown examples of belief persistence in 5, when weak evidence fails to change strongly held previous beliefs. But belief revision is also possible, since sufficiently strong new evidence or beliefs could knock one out of the current attractor basin. Consider Figure 7, where a very strong anti-deterrence study is presented. If the other studies get updated first, the strength of the new anti-deterrence study will flip their belief states, pushing the system into the attractor basin for a new anti-death penalty equilibrium.[7]

Figure 7 also includes an alternate version with an intermediate belief in whether deterrence works. The dynamics for these purposes are basically similar. However, there is a clear two step process in belief revision: first the agent changes his belief in whether deterrence works, then later will reassess his now dissonant old beliefs in the other studies.

Second, cognitive consistency biases occur only as a byproduct of *processing over time*. If a person is not trying to reconcile contradictions, or is not attending to them, they may persist. Actual experienced consonance is a function of both belief states and the distribution of attention on them. This implies that further deliberation and assessment of evidence could entrench a self-consistent worldview, which would manifest itself as evidence assimilation biases. Also, if a person does not have enough time to process and explain away destabilizing information — say, a barrage of new evidence or a moment of lucidity — that may allow new dissonant nodes to accumulate and then flip around the old beliefs. Finally, the rational case is also possible: given sufficient time and motivation, deliberative or intuitive processing

[7] However, if the new study gets updated first, it will be strongly revised to be disbelieved. This may be realistic: if you first pause to assess your old beliefs instead of explaining the new belief, that may be an important event to allow belief revision instead of persistence. But if we want to model incontrovertibly strong evidence, we can introduce for each node a bias_i constant in the activation sum, so $a_i = \text{bias}_i + \sum_j w_{ij} x_j$.

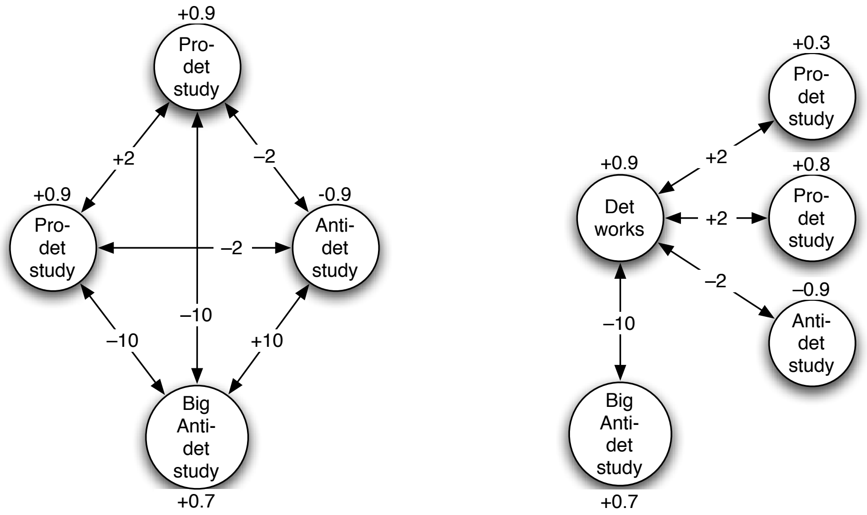


Figure 7: Strongly dissonant new evidence destabilizes the pro-death penalty equilibrium, prompting belief revision.

could find that in response to new contradictory information it is better to revise previous beliefs rather than explain away the new information.

Furthermore, this model can explain *group polarization.* The consistency network is only guaranteed to *locally* maximize consonance. The networks here all have two stable configurations: (1) belief in pro-deterrence and disbelief in anti-deterrence studies, and (2) vice-versa. If people are in different local maxima, a commonly viewed piece of evidence could be insufficient to knock them into the same attractor basin; instead, different prior beliefs imply different directions of consonance hill climbing, thus different interpretations of evidence.

Finally, this model may explain what Baron (2000) calls *belief overkill*: when people tend to believe all the good arguments are on one side, causing them to hold overly consistent views. For example, in another capital punishment attitude study, it was found that CP opponents tend to believe both that capital punishment is immoral, and that capital punishment is ineffective at deterring crime, while proponents tended to believe the opposite on both accounts. People who believed one but not the other were rare, though this should be logically possible. (Ellsworth and Ross, 1983). This is easily viewed in terms of a consistency network. “CP Moral” and “Deterrence Works” are not directly connected, or are weakly connected, but each is positively connected to a node “CP Good.” Believing one but not the other has to be a dissonant state because of the connections to “CP Good”; if the state that disagrees with “CP Good” is selected for updating, it will flip to conform.

4.3 Rational consistency?

Cognitive consistency networks are interesting because of their system dynamics of local maxima and processing over time. It seems like they can assimilate evidence in a way biased toward previously held beliefs. But it is not obvious why this is necessary, or why one could not simply accept mixed evidence and hold an indifferent opinion. To judge whether bias or correct information processing is taking place, we need to carefully formulate the question in terms of a well-specified probabilistic inference problem. To find the most likely explanation, we implement again a consistency search — or, in probabilistic terms, a maximum (log)-likelihood search. The following section finds clear-cut assimilation bias operating analogously to the properties observed above.

5 Bias in approximate Bayesian agents

In this section I argue that consistency search can be seen as an approximate Bayesian inference mechanism. Since it has problems getting stuck in local maxima, its results can be substantially different from exact Bayesian inference. In fact, I illustrate that consistency search can result in belief persistence bias, defined as a systematic deviation from Bayesian inference toward one’s prior beliefs.

5.1 Introduction to Bayesian inference

In the Bayesian view of probability theory, a probability is a degree of belief in a proposition. The laws of standard probability are a rational method of reasoning under uncertainty. At least two justifications exist. (1) The Cox axioms prescribe a few axiomatic desiderata for reasoning about degrees of belief, such as transitivity. It can be shown that any system of reasoning satisfying these axioms can be mapped to standard probability theory (Cox, 1961; Jaynes, 2003). (2) The “Dutch book” argument shows that a decision-maker that does not follow probability theory for its beliefs will accept losing gambles. Thus probability theory ensures optimal performance. (de Finetti, 1970; Savage, 1954). There is, of course, much more to these accounts, as well as many other approaches to the foundations of probability theory; see Suppes (2001) and Hajek (2003) for reviews.

Reassuringly, these different approaches for reasoning under uncertainty all imply the standard definition of conditional probability,

$$P(A|B) = P(AB) \ / \ P(B)$$

and standard identities such as $P(\bar{A}) = 1 - P(A)$.

We will extensively use two rules of probability theory. The first is the law of total probability, in joint and conditional versions:

$$P(A) = \sum_{b \in \mathrm{dom}(B)} P(A, b) \tag{1}$$

$$P(A) = \sum_{b \in \mathrm{dom}(B)} P(A|b) \ P(b) \tag{2}$$

where dom(B) is the set of possible values a discrete B can attain. The conditional version has a nice intuitive form: the probability of A is the weighted average of its likelihood across all different scenarios of B, weighted by the probability of each scenario.

The second rule is the celebrated Bayes' rule to flip a conditional:

$$P(H|D) = P(H) \ P(D|H) \ / \ P(D)$$

Bayes' rule is useful if we want to compute an updated belief in some hypothesis H after learning data D, and we know the likelihood of data given the hypothesis $(D|H)$. Bayes rule says to multiply the prior $P(H)$ by the likelihood $P(D|H)$, and normalize by $1/P(D)$. When viewing the posterior $P(H|D)$ as a function of H, the term $1/P(D)$ is just a constant, and it drops out in many contexts we examine. One such context is the odds-ratio form of Bayes' rule, obtained by dividing $P(H|D)$ by its negation $P(\bar{H}|D)$:

$$\begin{array}{ccccc} \dfrac{P(H|D)}{P(\bar{H}|D)} & = & \dfrac{P(H)}{P(\bar{H})} & \times & \dfrac{P(D|H)}{P(D|\bar{H})} \\ \underbrace{O(H|D)}_{\text{Posterior odds}} & = & \underbrace{O(H)}_{\text{Prior odds}} & \times & \underbrace{\dfrac{P(D|H)}{P(D|\bar{H})}}_{\text{Likelihood ratio}} \end{array}$$

The odds form $O(.)$ of a probability $P(.)$ is just $\frac{P(.)}{1-P(.)}$; thus the probability of 1/2 is 1:1 odds, probability 3/5 is 3:2 odds, and probability 1/10 is 1:9 odds, etc. This is also known as the racetrack or betting odds. The posterior odds can be seen as a comparison between the updated belief that H is true, versus the updated belief that H is false.

The odds form of Bayes' rule makes it very clear that to update one's belief in H upon learning D, you simply multiply your prior odds-belief $O(H)$ by the likelihood ratio. If D is more likely given H than it is given $\bar{H}$, then the likelihood ratio is high, and the posterior odds increase. If D is less likely under H than $\bar{H}$, then the likelihood ratio is less than one, so the posterior decreases. If the data is equally likely under either scenario, then the data is completely non-informative with a likelihood ratio of 1, so the posterior does not change.

The log-odds form of Bayes' rule makes its incremental nature very apparent, because Bayes updates simply proceed by adding the likelihood "weight" of a piece of evidence:

$$\log O(H|D) = \log O(H) + \log \frac{P(D|H)}{P(D|\bar{H})}$$

Thus you move closer to the data likelihood, but moderated by your prior belief. If your prior belief in H is non-informative, i.e. completely neutral (for a binary hypothesis, $P(H) = 1/2$, $O(H) = 1$, $\log O(H) = 0$), then your posterior is completely determined by the likelihood. A strong prior belief will move in the direction of the likelihood, but not be completely determined by it.

The prior belief is simply one's belief based on all evidence seen so far. The question of where priors come from without any evidence at all is a sticky topic in itself. Dodging the question, we note that there is always some background evidence Z informing any belief, so really all the previous Bayes' rules were shorthand for the form:

$$P(H|DZ) = P(H|Z)\ P(D|HZ)\ /\ P(D|Z)$$

where all probabilities are conditioned on a background Z. This background-conditional Bayes' rule will also be used in this section.

Final miscellaneous definitions: we say A and B are *marginally independent* if $P(AB) = P(A)\ P(B)$. A and B are *conditionally independent given* Z if $P(AB|Z) = P(A|Z)\ P(B|Z)$. Note that it can sometimes be useful to loosely interpret probabilistic independence as causal independence: if variables are independent, they do not cause or affect one another.

Good introductions to Bayesian probability and statistics include Lee (1997) and chapters 2 and 3 of MacKay (2003); see also Tom Griffiths' reading list on Bayesian methods for cognitive science.[8]

[8] Currently available at `http://cocosci.berkeley.edu/tom/bayes.html`

5.2 Problem #1: inference from simple evidence

An unknown binary hypothesis H may be in one of two states, $+1$ or -1, which we denote h^+ or h^-. Thus we abbreviate $P(H{=}{+}1)$ as $P(h^+)$. An agent observes a sequence of binary signals $D_i \in \{d_i^+, d_i^-\}$, which are identically and independently distributed (i.i.d.). That means two things. (1) They are conditionally independent given a particular value of H. A causal way to think of this conditional independence is that the hidden hypothesis sends out visible signals to the agent. And, (2) each D_i has the same distribution given H. Specifically, each signal corresponds with the true state of the hypothesis at probability θ:

$$\begin{array}{ccc} P(d_i^+ \mid h^+) = \theta & \text{and} & P(d_i^- \mid h^+) = 1-\theta \\ P(d_i^- \mid h^-) = \theta & \text{and} & P(d_i^+ \mid h^-) = 1-\theta \end{array}$$

Upon receiving each new signal, a Bayesian agent updates her beliefs via Bayes' rule. In odds form, the posterior belief in h^+ versus h^- after learning from one signal is

$$O(h^+ \mid d_1) = O(h^+)\ \frac{P(d_1|h^+)}{P(d_1|h^-)}$$

If a piece of evidence is positive, the likelihood ratio is $\frac{\theta}{1-\theta}$. A negative piece of evidence has a likelihood ratio of $\frac{1-\theta}{\theta}$. Assume $\theta > 0.5$, so positive evidence makes the h^+ more likely, and negative evidence makes h^- more likely.

For the second signal, Bayes' rule prescribes

$$O(h^+ \mid d_1, d_2) = O(h^+ \mid d_1)\ \frac{P(d_2|h^+, d_1)}{P(d_2|h^-, d_1)}$$

but since signals are conditionally independent given H,

$$= O(h^+ \mid d_1)\ \frac{P(d_2|h^+)}{P(d_2|h^-)}$$

Thus for every piece of new evidence d_t, the agent multiplies her current belief by the new likelihood ratio of d_t.

$$\begin{aligned} O(h^+ \mid d_1..d_{t-1}, d_t) &= O(h^+ \mid d_1..d_{t-1})\ \frac{P(d_t \mid h^+)}{P(d_t \mid h^-)} && (3) \\ &= O(h^+ \mid d_1..d_{t-1}) \left(\frac{\theta}{1-\theta}\right)^{d_t} && (4) \\ &= O(h^+) \prod_{i=1..t} \left(\frac{\theta}{1-\theta}\right)^{d_i} \quad \text{by conditional indep.} && (5) \end{aligned}$$

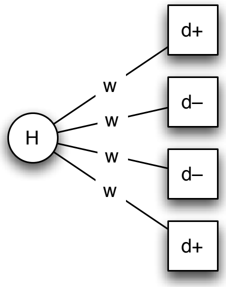


Figure 8: Hopfield network that calculates a Bayesian point estimate of H (e.g., MAP on the linear threshold rule). Boxed nodes are observed variables; circles are hidden.

5.3 Sidenote: Hopfield constraint satisfaction as probabilistic inference

We can represent the agent's reasoning as a very simple Hopfield network. Each node represents a variable. There is only one updatable node (hidden variable) H. Each node D_i is an observed variable, with its state clamped to $+1$ or -1, and is connected only to H. All H-D_i links have uniform weight w. Note the input activation is the same as the posterior log-odds of h^+ if $w = \log \frac{\theta}{1-\theta}$, so $wd_i^+ = \log \frac{\theta}{1-\theta}$ and $wd_i^- = \log \frac{1-\theta}{\theta}$:

$$
\begin{aligned}
a_H &= \text{bias}_H + \sum_i w d_i \\
\log O(h^+ \mid \vec{d}) &= \log O(h^+) + \sum_i \log \frac{P(d_i \mid h^+)}{P(d_i \mid h^-)}
\end{aligned}
$$

If $a_H > 0$, the posterior belief in H is tilted towards h^+. If we had to pick one value of H to believe in, the most likely value is h^+. Thus the linear threshold update rule $x_H := \text{sign}(a_H)$ picks the maximum posterior value of H, called the MAP estimate: $x_H = \arg\max_h P(h \mid \vec{d})$. ("MAP" helpfully stands for "maximum a posteriori.") A version of this Bayesian view of input activation terms is documented in Hinton and Sejnowski (1983) and reviewed in McClelland (1998); see also Jordan (1995).

This is only a small illustration of the relationship between probability and neural networks. The probabilistic semantics of weights in neural networks becomes more complex for symmetric networks and multilayered networks.

5.4 Behavioral signal misperception model

Rabin and Schrag (1999) construct a behavioral model of evidence assimilation bias through signal misperception. A biased agent does Bayes updates, except when her prior is tilted in one direction; in that case, she may misperceive a contradictory signal as actually supporting the prior. The biased agent perceives binary signals $\delta_i \in \{\delta^+, \delta^-\}$ that have a q probability of being mistaken in the biased case, but perfectly correspond when confirming the agent's prior bias. For when the agent's prior is tilted toward h^+ (that is, $P(h^+|d_1..d_{t-1}) > 0.5$):

$$\begin{array}{lll} P(\delta_t^+ \mid d_t^+) = 1 & \text{and} & P(\delta_t^- \mid d_t^+) = 0 \\ P(\delta_t^- \mid d_t^-) = 1 - q & \text{and} & P(\delta_t^- \mid d_t^-) = q \end{array}$$

Thus the sequence of perceived signals δ_t are not conditionally independent given H, so the agent's beliefs depend on the order of true d_t signals. Rabin and Schrag go on to demonstrate various facts about biased agents, such as that they tend to be overconfident, and that an infinite sequence of signals does not guarantee they will arrive at the correct conclusion. These are interesting findings, but I am interested in an *explanation* of this bias.

This model does suggest a simple definition of evidence assimilation bias as a deviation from Bayesian updating. Since the evidence interpretation problem I present is multivariate, Rabin and Schrag's setup cannot be directly applied. However, if negative evidence against the currently favored hypothesis fails to be counted correctly, we say bias exists.

5.5 Problem #2: Inference from interpretable evidence

As before, a sequence of conditionally independent pieces of evidence D_i correspond with the value of H at probability $\theta > 0.5$. However, each piece of evidence could potentially be explained away by a mediating variable $B_i \in \{b^Y, b^N\}$. If B_i is Y, the evidence is believed/credible, so it does correlate at probability θ with H. But b^N explains away the evidence — the study was conducted poorly, the source is not credible, we misread/misheard it, etc. We should not believe the evidence. In that case, d_i^+ occurs at uniform probability 1/2: it is completely non-informative about H. For convenience, we will often call B_i "credibility," though it could refer to number of factors affecting the evaluation of the evidence. Like H, the B_i

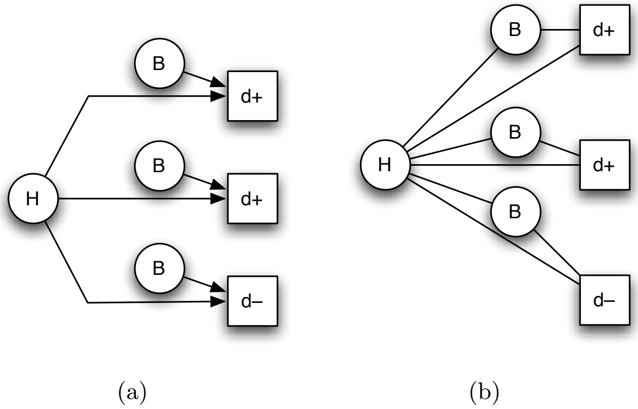
are hidden: it is impossible to know for sure whether a piece of evidence can be trusted, but you can make judgments from all the evidence available. Finally, their influence is local: each B_i only affects that D_i.

Specifically, each conditional distribution is

$$\begin{aligned} P(d_i^+ \mid h^+ b_i^Y) &= \theta \\ P(d_i^+ \mid h^- b_i^Y) &= 1-\theta \\ P(d_i^+ \mid h^+ b_i^N) &= 1/2 \\ P(d_i^+ \mid h^- b_i^N) &= 1/2 \end{aligned}$$

Given the observed evidence $\vec{d}$, one problem is to infer a distribution over the hidden variables $(H, B_1, ..., B_n)$. Another problem is to compute a point estimate $(h, b_1..b_n)$, e.g. the MAP estimate. To do either of these tasks, we need to work with a representation of the joint distribution $P(H, \vec{B}, \vec{D})$. Given the dependency relationships between the variables, we can compactly and illustratively represent the distribution as a directed probabilistic graphical model, more often called a *Bayesian network* (Koller and Friedman, 2006; Murphy, 1998; Pearl, 1988). A BN has two components: (1) a directed acyclic graph representing qualitative dependencies between variables, and (2) a set of conditional distributions for each child given on its parents.

Figure 9: Directed and undirected graphs, i.e. Bayesian network and Markov random field, for the evidence interpretation problem. (a) The BN's directed links indicate direct conditional dependence or causality. (b) The MRF's undirected links document all possible direct probabilistic influences, including explaining away.



(a) (b)

The DAG is shown in Figure 9. Arrows can sometimes be interpreted as causality; we only take them to denote local conditional dependencies. The graph's semantics are defined as the *local Markov property*: a node is conditionally independent of its non-descendants given its parents. This allows a BN's joint distribution over nodes $\vec{X}$ to be factorized as the product of local conditional distributions $P(\vec{X}) = \prod_i P(X_i | \mathrm{Pa}(X_i))$, where $\mathrm{Pa}(.)$ denotes a node's immediate parents.

With the notation $A \perp B \mid Z$ meaning A and B are conditionally independent given Z, this graph encodes a number of conditional independencies, including:

- $H \perp B_i$ and $B_i \perp B_j$: the hypothesis and credibilities are a priori (marginally) independent.

- $B_i D_i \perp B_j D_j \mid H$: credibility/evidence joint pairs occur independently given a hypothesis.

This problem is interesting for the relationships that are *not* independent. Note the v-structures: $H \to D_i \leftarrow B_i$. Since H and B_i are root nodes, they are marginally independent. However, they are conditionally dependent upon observing d_i. For example, a positive d_i^+ causes a correlation between its credibility B_i and one's belief in h^+: if we have seen positive evidence, it is unlikely the evidence is credible but the hypothesis is false, or vice-versa. An intuitive way to think of this is that probabilistic influence flows in the direction of the links, but is "blocked off" by an observed node. Thus the influence from B_i, flows down to the blocked-off d_i, then flows up out of the v-structure to H. (This intutition is formalized in the Bayes-Ball algorithm (Shachter, 1998).)[9]

Thus problematic dependencies include:

- $H \not\perp B_i \mid D_i$: Your belief in H affects your assessment of d_i's credibility.

- $B_i \not\perp B_j \mid D_i, D_j$: The credibilities of different pieces of evidence affect one another (because they affect your belief in H).

[9] This model implements "explaining away" through the B_i variables, but the term "explaining away" often refers to an endogenous effect arising in disjunction-like v-structures. To use an oft-repeated example: R is whether it rained last night, S is whether the sprinkler was left on, and W is whether the lawn is now wet. (Network: $R \to W \leftarrow S$.) R and S are conditionally dependent given W. For example, given the grass is wet, learning that it rained last night decreases your belief the sprinkler was left on, because the rain explains away the fact the lawn is wet.

So all hidden variables $H, B_1..B_n$ are conditionally dependent given the observations $\vec{d}$. This is the crux of the problem: the impact of observed evidence d_i on H depends on how you interpret it (B_i), but how you interpret it depends on H and thus the interpretation of other evidence. Simon et al. (2004) seem to have something like this in mind when they declare humans do not use "unidirectional" reasoning; but fortunately, it *is* possible to deploy Bayesian reasoning to solve this problem.

Finally, to complete the specification of the distribution we use the conditional distributions for each node, given its immediate parents. These $P(H)$, $P(B_i)$, and $P(D_i|HB_i)$ tables together define a joint probability distribution over all variables $H, \vec{B}, \vec{D}$:

$$P(H, \vec{B}, \vec{D}) = P(H) \prod_i P(B_i) \prod_i P(D_i|HB_i)$$

5.6 Exact inference solution

We want to compare optimal Bayesian behavior to approximate algorithms, so we first we compute the hidden distributions via exact Bayesian inference. We can solve the problematic dependencies by marginalizing out root nodes as needed; this procedure is more generally formalized as the variable elimination and message passing algorithms (Pearl, 1988).

Using the conditional independence assumptions, it is convenient to calculate the posterior of H:

$$\begin{aligned}
P(h^+|\vec{d}) &= P(h^+)\ P(\vec{d}|h^+)\ /\ P(\vec{d}) \text{ by Bayes' rule} && (6)\\
&= P(h^+)\ \frac{1}{P(\vec{d})} \prod_i P(d_i|h^+) \text{ by local Markov property} && (7)\\
&= P(h^+)\ \frac{1}{P(\vec{d})} \prod_i \sum_{b_i \in \{Y,N\}} P(d_i|h^+, b_i)\ P(b_i) && (8)\\
O(h^+|\vec{d}) &= O(h^+) \prod_i \frac{\sum_{b_i} P(d_i|h^+, b_i)\ P(b_i)}{\sum_{b_i} P(d_i|h^-, b_i)\ P(b_i)} && (9)\\
O(h^+|\vec{d}) &= O(h^+) \prod_i \left(\frac{\theta\ P(b_i^Y)\ +\ \frac{1}{2}P(b_i^N)}{(1-\theta)\ P(b_i^Y)\ +\ \frac{1}{2}P(b_i^N)} \right)^{d_i} && (10)
\end{aligned}$$

Contrast to the exact inference solution for the problem without credibilities (Equation 5), where the likelihood ratio is $\left(\frac{\theta}{1-\theta}\right)^{d_i}$. By summing out each prior $P(B_i)$, it is apparent that each evidence's likelihood $P(d_i|h^+)$ has been moderated towards 1/2 by the chance that b_i is N. That is, this solution takes into account the chance the evidence is not reliable.

5.7 Bias in local MAP search

A Bayesian updater computes and re-computes the distribution over all hidden parameters $(H, B_1..B_n)$. This doesn't require just a marginal distribution over each parameter, but rather a distribution over all *combinations* of parameters. The state space of the hidden vector is the cross-product $\text{dom}(H) \times \prod_i \text{dom}(B_i)$; for our binary variables, this is size 2^{n+1}. For this particular problem with simple uniform conditional probabilities, there may be a simpler representation with smaller storage requirements; however, it seems that in general, belief networks among thousands of beliefs with complex interconnections, storing an entire joint distribution grows exponentially in the number of hidden variables.

A less memory-intensive problem is to store only one particular instantiation of the hidden parameters — that is, a single belief vector $(h, b_1..b_n)$. This assumes people remember only categorically: evidence i was either credible or not, the hypothesis is either true or not, etc.[10] The most likely explanation for $\vec{d}$ is the MAP estimate

$$\arg\max_{h,\vec{b}} P(h, \vec{b}|\vec{d})$$

The belief revision problem is to update this vector as new evidence comes in. A simple local MAP search algorithm to accomplish this is coordinate ascent on the posterior: iterate through the elements of $(h, b_1..b_n)$ and independently optimize each in turn, and continue iterating until no more such updates can be made. Note that optimizing the posterior with regards to $h, \vec{b}$ is the same as optimizing the joint with regard to those two variables:

$$\begin{aligned} & \arg\max_{h,\vec{b}} P(h, \vec{b}|\vec{d}) = \arg\max_{h,\vec{b}} P(h, \vec{b}|\vec{d})P(\vec{d}) = \arg\max_{h,\vec{b}} P(h, \vec{b}, \vec{d}) \\ = \; & \arg\max_{h,\vec{b}} P(h) \prod_i P(b_i)P(d|h, b_i) \end{aligned}$$

The only term that is a function of b_i is $P(b_i)P(d_i|h, b_i)$. The only terms that are function of h are $P(h)\prod_i P(d|h, b_i)$. Thus to individually optimize each one, we can drop out all other terms:

$$b_i \;:=\; \arg\max_{b_i} P(b_i)P(d_i|h, b_i) \tag{11}$$

$$h \;:=\; \arg\max_{h} P(h) \prod_i P(d_i|h, b_i) \tag{12}$$

[10] Even if people can store distributions over each of these, a distribution over joint combinations seem unreasonable; Section 5.8 argues per-variable distributions would encounter the same biases as found in local MAP search.

This demonstrates the local properties of a graphical network. Each b_i only has to be optimized with regards to the local triad of (h, b_i, d_i). Similarly, h can ignore the b_i root prior distributions.

This can also be seen when viewing the MAP inference problem as a soft constraint satisfaction problem, a la the Hopfield networks from Section 4. The log probability is the sum of single and three-way binary constraints:

$$\begin{aligned} \log P(h, \vec{b}, \vec{d}) &= \log P(h) + \sum_i \log[P(b_i)\ P(d_i|h, b_i)] && (13) \\ -E(h, \vec{b}, \vec{d}) &= \phi_0(h) + \sum_i \phi_i(h, b_i, d_i) && (14) \end{aligned}$$

The consonance function is the log of the joint probability. To optimize b_i, we only need to look at ϕ_i, the only constraint involving b_i. The undirected graph representing these dependencies is shown in Figure 9, in which a connection between two nodes means there exists a constraint function ϕ involving both of them. (The representation of a probability distribution with an undirected graph and these constraint functions ("log clique potentials") is called a Markov random field, a.k.a Markov network.) This suggests that all of the observations from Section 4 on Hopfield networks could also apply to MAP search on probabilistic graphical networks.

Consider an agent updating a local MAP explanation with every new piece of evidence. Assume all b_i and h have uniform priors: $P(b_i^+) = 0.5$, $P(h^+) = 0.5$. Thus the priors $P(h)$ and $P(b_i)$ drop out of the argmax Equations 11 and 12. The update rules are very simple:

$$b_i := \begin{cases} Y \text{ if } h \text{ and } d_i \text{ agree} \\ N \text{ if they disagree} \end{cases} \qquad (15)$$

(Since $\theta > 1/2$: agreeable evidence is more likely to be credible/true.) (16)

$$h := \begin{cases} h^+ & \text{if there are are more } (Y, +) \text{ than } (Y, -) \text{ pairs} \\ h^- & \text{if there are are more } (Y, -) \text{ than } (Y, +) \text{ pairs} \\ h & \text{if there are the same number} \end{cases} \qquad (17)$$

(Since only credible evidence is relevant; break ties with no change.) (18)

The agent's belief state is order-dependent and displays assimilation bias effects. Consider the following example.

Example An agent receives a sequence of positive and negative signals d_i, each with an initially positive evaluation b_i^Y. Upon receiving each piece of evidence, the agent searches for a local MAP explanation via the coordinate-ascent rules described above.

With no evidence at $t = 0$ the initial assignment is (h^+). Consider the sequence of data $\vec{d} = (+, +, -, -, -)$. At $t = 1$ a new evidence/credibility pair $(Y, +)$ is added. The belief vector (h^+, b_1^Y) is stable since neither variable will change when selected for an update. b_1^Y is locally stable: since h and d_1 agree, it is more likely that the evidence is credible than not. h^+ is also locally stable, since there is 1 piece of credible positive evidence and no pieces of credible negative evidence.

$t = 2$ sees the addition of another credible, positive evidence pair $(Y, +)$, which results in the similarly stable belief (h^+, b_1^Y, b_2^Y). $t = 3$ adds a credible negative pair $(Y, -)$. This is not stable. The local MAP search tries to optimize all variables in turn, but only b_3 will be revised. It flips to N to explain away the negative evidence. H stays positive since there are more positive than negative instances. This is reasonable, since the negative evidence is too weak at this point to revise that belief. The better explanation for the evidence is that it is not credible. Finally, if b_1 or b_2 are assessed, they will not flip either, since the best explanation for positive evidence is the that they're credible, given that H is true.

The process continues like so:

$t = 2$	$h^+\ (Y,+)(Y,+)$	
$t = 3$	$h^+\ (Y,+)(Y,+)(Y,-)$	receive new data
	$h^+\ (Y,+)(Y,+)(N,-)$	revise b_3
$t = 4$	$h^+\ (Y,+)(Y,+)(N,-)(Y,-)$	receive new data
	$h^+\ (Y,+)(Y,+)(N,-)(N,-)$	revise b_4
$t = 5$	$h^+\ (Y,+)(Y,+)(N,-)(N,-)(Y,-)$	receive new data
	$h^+\ (Y,+)(Y,+)(N,-)(N,-)(N,-)$	revise b_5

If an agent continues receiving d^- signals and (1) local MAP updates only one node at a time, and (2) does all possible updates each time when given evidence, then the agent will suffer from evidence assimilation bias. The first few positive signals and the prior h^+ set up the agent for evidence assimilation bias: later negative signals will be explained away, even though there are more negative signals than positive ones.

This is a bias because the agent has the wrong answer: it is stuck at a local maximum on the posterior surface. If the agent was able to consider the entire vector of credibilities $\vec{b}$, it could update to a higher posterior estimate $(b_1^N, b_2^N, b_3^Y, b_4^Y, b_5^Y ...)$ where the positive signals are uncredible and the negative signals are credible.

To more formally analyze this behavior, we note several facts.

Proposition Given the coordinate ascent procedure, at $t \geq 1$ there are just two local maxima for the belief vector $\vec{x} = (h, b_1..b_t)$, denoted $\vec{x}^+$ and $\vec{x}^-$:

$$\vec{x}^+ = \begin{cases} h^+ \\ b_i^Y & \text{for } i \text{ where } d_i = +1 \\ b_i^N & \text{for } i \text{ where } d_i = -1 \end{cases} \tag{19}$$

$$\vec{x}^- = \begin{cases} h^- \\ b_i^N & \text{for } i \text{ where } d_i = +1 \\ b_i^Y & \text{for } i \text{ where } d_i = -1 \end{cases} \tag{20}$$

$$\tag{21}$$

That is, either the agent believes the hypothesis and only positive evidence, or the agent disbelieves the hypothesis and believes only the negative evidence.

Proof: First note that $\vec{x}^+$ and $\vec{x}^-$ are both local maxima. h cannot revise because all the credible evidence favors its current state. None of the b_i's can revise because they all conform to the current hypothesis; switching any b_i to a nonconformant status would decrease likelihood.

Furthermore, no other $\vec{x}$ vectors are local maxima. Say $\vec{x}$ has h^+ but there is a nonconforming evidence belief, either a positive evaluation of negative evidence, or negative evaluation of positive evidence. (1) Say some b_i^Y has a negative d_i^-. This $\vec{x}$ is not stable, because b_i would update to b_i^N. (Furthermore, it is possible that h could flip to h^- if there was was sufficient credible negative evidence.) The other possible nonconformant evidence belief is (2) some b_j^N is on a positive d_j^+. This b_j is similarly unstable. An analogous argument holds for $\vec{x}$ with h^- and nonconforming evidence beliefs.

Proposition Assume uniform $P(b_i)$ priors. After n signals, with n^+ positive signals and n^- negative signals, the globally optimal MAP $\vec{x}$ is $\vec{x}^+$ if

$$(n^+ - n^-) \log 2\theta > \log O(h^-)$$

That is, sufficiently many positive signals cause $\vec{x}^+$ to be better, since $\theta > 0.5$ so $\log 2\theta$ is positive.

Proof: note that the joint probability of $\vec{x}^+$ and the evidence can be broken down into the h^+ prior times the probability of credible positive evidence and noncredible negative evidence:

$$P(\vec{x}^+, \vec{d}) = P(h^+) \prod_i P(b_i) P(d_i | b_i, h^+)$$

$$P(\vec{x}^+, \vec{d}) = P(h^+) \prod_{i \text{ where } d_i^+} P(b_i^Y) P(d_i^+ | b_i^Y, h^+) \prod_{j \text{ where } d_j^-} P(b_i^N) P(d_i^- | b_i^N, h^+)$$

Next: $\vec{x}^+$ is more likely than $\vec{x}^-$ if

$$1 < \frac{P(\vec{x}^+ | \vec{d})}{P(\vec{x}^- | \vec{d})}$$

$$1 < \frac{P(\vec{x}^+, \vec{d})}{P(\vec{x}^-, \vec{d})}$$

$$1 < O(h^+) \prod_{i \text{ where } d_i^+} O(b_i^Y) \frac{P(d_i^+ | b_i^Y, h^+)}{P(d_i^+ | b_i^N, h^-)} \prod_{j \text{ where } d_j^-} O(b_j^N) \frac{P(d_j^- | b_j^N, h^+)}{P(d_j^- | b_j^Y, h^-)}$$

$$1 < O(h^+) \prod_{i \text{ where } d_i^+} \frac{P(d_i^+ | b_i^Y, h^+)}{P(d_i^+ | b_i^N, h^-)} \prod_{j \text{ where } d_j^-} \frac{P(d_j^- | b_j^N, h^+)}{P(d_j^- | b_j^Y, h^-)} \text{ since uniform priors on } b_i$$

$$1 < O(h^+) \prod_{i \text{ where } d_i^+} \frac{\theta}{1/2} \prod_{j \text{ where } d_j^-} \frac{1/2}{\theta}$$

$$1 < O(h^+)(2\theta)^{(n^+ - n^-)}$$

$$0 < \log O(h^+) + (n^+ - n^-) \log 2\theta$$

The key result of this section is that *belief persistence bias is often possible*:

Proposition (Possibility of persistence bias): Assume uniform $P(h), P(b_i)$. Also assume the agent locally optimizes after learning each individual piece of evidence through coordinate ascent, iterating through the $\vec{x}$ vector in the order $(h, b_1..b_t)$.

Then, given a set of $n > 1$ signals favoring the positive configuration $\vec{x}^+$ with both positive and negative signals, there exists an order of those signals such that after learning all of them, the agent mistakenly believes $\vec{x}^-$. That is, he takes positive evidence as noncredible, and negative evidence as credible, i.e., exhibits belief persistence bias.

Proof: Since the agent locally maximizes at every timestep, it turns out the first signal completely determines future beliefs. If the negative evidence comes first, $t = 1$ initially has $(h^+, (Y, -))$ By updating h first, the agent settles on $\vec{x}^+ = (h^-, Y)$. Any future positive evidence will fail to revise h back to h^+, so will instead be taken as noncredible.

Thus the coordinate ascent search procedure has incredibly strong bias. A sketch of smarter search procedures follows. They can have less bias, but are still susceptible in various ways.

One way out of this bias is less local search. Say a more cognitively empowered agent selects k variables at a time to optimize, and considers all possible subsets size k as candidates for updating. Then the sequence $\vec{d} = (+, +, -, -, -)$ yielding the erroneous $\vec{x}^+ = (h^+, Y, Y, N, N, N)$ could get properly flipped to the global maximum $\vec{x}^- = (h^-, N, N, Y, Y, Y)$ with just $k = 3$. But for a fixed value of k, just $k + 2$ prior instances makes it impossible to update out of the wrong local maximum.

A second way out of the bias is less greedy search. Currently the agent locally maximizes the MAP estimate every time a new piece of evidence is presented. An agent that never optimizes b_i's, leaving them at b^Y, will make the correct MAP estimate for H if all b_i have the same prior, though he may have many errors for the b_i values. An agent sacrifices the most likely explanations for the evidence in return for the best explanations on h.

Or say an agent wants the most likely explanation for b_3 — say, d_3 has an important competing hypothesis to test. The agent should use b^Y for all other beliefs, updating h accordingly, then assign b^Y to conform between h and d_3. The agent then has a decent evaluation of h involving almost all the evidence, a poor evaluation of the evidence where much noncredible evidence is believed, but will have a good estimate of b_3.

Another alternative search procedure is stochastic exploration, in which an agent may sometimes update a variable to the less likely state. This allows exploration out of local maxima. This represents an open-minded (or perhaps non-opinionated) agent willing to live with mixed evidence, and therefore less coherent explanations for the world; his beliefs have less likelihood than the beliefs of the local searcher, but may be closer to the maximally likely state of beliefs (and, hopefully, closer to the true state of the world.)

One form of stochastic exploration is Gibbs sampling: Given a current belief vector $\vec{x}$, to update some element x_i, randomly choose $x_i := +1$ at probability $P(x_i^+ | \vec{x}_{-i}, \vec{d})$, else choose $x := -1$.[11] That is, instead of choosing

[11] This is just the sampling component of the Gibbs sampling algorithm; the actual use

the maximizing value of x_i as in coordinate ascent, it probability matches instead: if x_i^+ is much more likely than x_i^-, it probably chooses that, though there is a chance to take the less likely option.

Thus if the agent is at the local maximum $\vec{x}^+$, when updating h it will probably keep to h^+, though there is a small chance of exploring to h^-. If that happened, the beliefs b_i are more likely to flip to conform to disbelieving the hypothesis. An agent that used stochastic exploration could explore to $\vec{x}^-$; with some sort of memory of the likelihood of previously seen $\vec{x}$ states, or a combination of occasional stochastic exploration with local maximization, the agent could find better local maxima.

However, there is still some amount of bias from the initial state. If the order of evidence put the agent's believed $\vec{x}$ in a suboptimal maximum, the agent needs to get lucky in order to explore out of that maximum.

5.8 Psychologically plausible inference mechanisms

There are two attractions of modeling human psychology via approximate Bayesian algorithms. Compared to optimal Bayesian models, this approach allows much more flexibility to model phenomena in judgment and social psychology, where findings of mistakes and inefficiencies are widespread. Compared to ad hoc algorithms, the advantage is clear semantics and rich interpretations for what people are doing. The advent of probabilistic graphical models, popularized in AI during the 90's, provides intuitive and sophisticated frameworks to understand many types of problems in a probabilistic fashion (Chater et al., 2006). Furthermore, the fast-growing literature in computational statistics and Bayesian machine learning suggest many potential algorithms as hypotheses to analyze for psychological plausibility and to test experimentally.

Starting with the MAP search already developed, note that many psychological factors can be interpreted as constraints or methods of MAP search. For example, the number of variables an agent can simultaneously optimize is bounded by its short-term memory capacity.

Where an agent searches may be guided by memory as well. Imperfect recall implies it is hard to revise interpretations of old evidence. For example, during a political campaign people update their beliefs about political candidates based on various evidence, but afterwards, can only remember

of the algorithm is to calculate the entire distribution of $\vec{x}$, since in the limit (Markov chain steady state) it visits different $\vec{x}$ states at frequencies proportional to their joint likelihood $P(\vec{x}|\vec{d})$ (MacKay, 2003; Geman and Geman, 1984). It seems unclear to what extent this is psychologically plausible.

their belief about the candidate, but not the evidence used to arrive at that conclusion (the "paradox of the forgetful voter"; see Lodge and Taber (2000)). That implies people can't revisit old b_i variables, suggesting belief persistence bias will occur when assimilating contradictory evidence in the future.

Human memory is associative: remembering certain beliefs or facts makes it easier to bring related items to mind, or may even retrieve them automatically. If a set of b_i variables have strong associations with one another, a person revising one may remember the others and be inclined to revise them as well. Furthermore, locally optimizing a specific belief should only depend on other beliefs that can be remembered at the time.

This has been posited as an explanation for "debriefing paradigm" experiments. Ross, Lepper, and Hubbard (1975) had participants assess whether supposed suicide notes were real or faked, and were then told whether they were correct or not. Participants that had positive feedback thought their skill at the task was high. After this was done, they were informed that actually the feedback had been completely random. Participants that had received positive feedback still believed they were relatively skilled at the task. The associative recall explanation is that hearing the positive feedback reminds participants of other salient facts, such as their own skill at empathy in other situations. Though the feedback evidence was neutralized, when participants assess their skill at the task, those now-present memories are used to form a positive assessment.[12]

For future work, memory is an interesting psychological mechanism to explore because it is relatively well-understood compared to certain other cognitive processes (e.g. problem solving). For example, Mullainathan (2002) extensively develops a model of a Bayesian updater with imperfect memory, that could potentially be applied to evidence interpretation problems.

Many other mechanisms can be usefully interpreted as mechanisms for the MAP search. Motivation and affect can be seen as exogenous factors that guide the search. Some theories, such as cognitive dissonance, in fact view them as the primary cause of consistency biases (Festinger, 1957; Kunda, 1990), though this paper shows that computational limitations are sufficient to cause assimilation bias. It may be useful to view affect and motivation as components of a Bayesian approximation algorithm.

As mentioned in Section 4, the environment can be seen as a force that

[12] In a similar vein, participants may construct causal explanations for fictional evidence, which remain as valid reasons for their revised belief after debriefing (Anderson, Lepper, and Ross, 1980).

mediates attention, guiding the search. For example, consider Figure 2 depicting the tension of believing Catholicism, Republicanism, and holding views on abortion and the death penalty. Any assignment to these four beliefs must violate one of the constraints. That is, likelihood of the belief vector (Cath-true, Repub-true, abortion-bad, death-penalty-good) is low because the joint pair (Catholicism-true, death-penalty-good) has a low consonance — the Catholic church opposes the death penalty. However, if the issue of abortion is much more commonly attended to than the issue of the death penalty — say, for exogenous reasons it is more often reported in the media — then the *experienced* likelihood is much higher, since the negative constraint is rarely noticed.

The plausibility of MAP search can also be improved. The motivation for the local MAP model in the previous section was that Bayes updating the entire distribution over all possible $(h, b_1..b_n)$ vectors was psychologically implausible due to the computational costs involved. Point estimates, such as a MAP vector, have just a linear storage requirement, and thus are computationally plausible.[13]

An immediate objection that was glossed over is that, while it is unrealistic for people to store degrees of belief over all combinations of beliefs, surely people can entertain shades of gray — e.g., store one degree of belief for each variable (or, for non-binary variables, one distribution per variable). For example, in the binary $(H, \vec{B})$ inference example, people may search for a different estimate, the expected vector $E[(H, B_1..B_n)] = (E[H], E[B_1]..E[B_n])$. Thus a person's beliefs consist of the average estimates for each variable. For the probabilistic interpretation of a Hopfield network, this formulation is embodied in the continuous sigmoid update rule, which computes conditional expectations $x_i := E[x_i|\vec{x}_{-i}, \vec{d}]$. This is still less information than a probability measure over all joint $\vec{x}$ combinations; in particular, an average value cannot adequately represent two extreme maxima such as $\vec{x}^+$ and $\vec{x}^-$, which are exact opposites of one another, located at different corners of the space. It would be interesting to investigate whether expectation updating exhibits bias in similar ways to MAP updating.

In general, combinations of computing locally optimal values for some hidden variables, and locally expected values for other hidden variables, are

[13] Note that with certain assumptions on the hidden distribution $\vec{X}$, it could be possible to efficiently store the distribution. If each X_i is real-valued and $\vec{X}$ is a multivariate Gaussian, then it only takes $n + n^2$ storage space to represent the mean vector and convariance matrix, which together completely describe the distribution. If, as was implicitly assumed in Section 5.5, the representation needs to handle inference for any arbitrary multivariate distribution, such storage savings do not exist.

forms of Expectation-Maximization algorithms (Neal and Hinton, 1998), or more generally, variational methods (Jordan et al., 1999), which converge to local optima but are not guaranteed to find the best solutions. In fact, these algorithms can be seen as optimizing an approximating distribution $Q(\vec{X}|\vec{d})$ that is a mathematically and computationally simpler representation than $P(\vec{X}|\vec{d})$.[14] For example, the continuous Hopfield rule $x_i := \tanh(\sum_j w_{ij} x_j)$ optimally approximates a "mean field" distribution $Q(.)$ that only considers the interaction of average values of the variables, instead of potentially complex interactions of different joint combinations of them (see MacKay chs. 33, 42 (2003), as well as Jordan et al.). This seems analogous to reasoning based on exemplars or typical cases. Finding and testing optimal behavioral properties of psychologically plausible approximate representations is an interesting possibility for future research.

[14] Specifically, a variational algorithm updates $\vec{X}$ to minimize the relative entropy — a sort of distance measure between probability distributions (Kullback and Leibler, 1951) — between $Q(\vec{X}|\vec{d})$ and $P(\vec{X}|\vec{d})$.

References

R. P. Abelson, E. Aronson, W. J. McGuire, T. M. Newcomb, M. J. Rosenberg, and P. H. Tannenbaum, editors. *Theories of Cognitive Consistency: A Sourcebook*. Rand McNally, Chicago, 1968.

C. A. Anderson, M. R. Lepper, and L. Ross. Perseverance of social theories: The role of explanation in the persistence of discredited information. 39: 1037–1049, 1980.

S. E. Asch. Forming impressions of personality. *Journal of Abnormal and Social Psychology*, 41:258–290, 1946.

J. Baron. *Thinking and Deciding*. Cambridge University Press, 3rd edition, 2000.

N. Chater, J. B. Tenenbaum, and A. Yuillie. Probabilistic models of cognition: Where next? *Trends in Cognitive Sciences*, 10(7):292–293, 2006.

J. Coooper and R. Fazio. A new look at dissonance theory. *Advances in experimental social psychology*, 17:229–266, 1984.

R. T. Cox. *The Algebra of Probable Inference*. Johns Hopkins University Press, Baltimore, MD, 1961.

B. de Finetti. *Theory of Probability*. Wiley and Sons, New York, 1970.

P. Domingos, S. Kok, H. Poon, M. Richardson, and P. Singla. Unifying logical and statistical AI. In *Proceedings of the Twenty-First National Conference on Artificial Intelligence (AAAI-06)*, 2006. Available online: http://www.cs.washington.edu/homes/pedrod/papers/aaai06c.pdf.

A. J. Elliot and P. G. Levine. On the motivational nature of cognitive dissonance as psychological discomfort. *Journal of Personality and Social Psychology*, 67(3):382–394, 1994.

P. C. Ellsworth and L. Ross. Public opinion and capital punishment: A close examination of the views of abolitionists and retentionists. *Crime and Delinquency*, 29:116–119, 1983.

L. Festinger. *A Theory of Cognitive Dissonance*. Stanford University Press, 1957.

S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 6, pages 721–741, 1984. Available online: `http://www.cis.jhu.edu/publications/papers_in_database/GEMAN/GemanPAMI84.pdf`.

H. B. Gerard and G. C. Mathewson. The effects of severity of initiation on liking for a group: A replication. *Journal of Experimental Social Psychology*, 2:278–287, 1966.

A. Hajek. Interpretations of probability. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. 2003. Available online: `http://plato.stanford.edu/archives/sum2003/entries/probability-interpret/`.

F. Heider. *The Psychology of Interpersonal Relations*. Wiley, New York, 1958.

G. E. Hinton and T. J. Sejnowski. Optimal perceptual inference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1983. Available online: `http://papers.cnl.salk.edu/PDFs/Optimal%20Perceptual%20Inference%201983-646.pdf`.

K. J. Holyoak and D. Simon. Birdirectional reasoning in decision making by constraint satisfaction. *Journal of Experimental Psychology: General*, 128(1):3–31, 1999.

J. J. Hopfield. Neural networks and physical systems with emergent collective computational abilities. In *Proceedings of the National Academy of Sciences*, 1982.

E. T. Jaynes. *Probability Theory: The Logic of Science*. Cambridge University Press, 2003. Earlier version online: `http://omega.albany.edu:8008/JaynesBook.html`.

M. I. Jordan. Why the logistic function? a tutorial discussion on probabilities and neural networks. 1995. Computational Cognitive Science Technical Report 9503. Available online: `http://www.cs.berkeley.edu/~jordan/papers/uai.ps.Z`.

M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37:183–233, 1999. Available online: `http://www.cs.berkeley.edu/~jordan/papers/variational-intro.ps.gz`.

D. Koller and N. Friedman. *Structured Probabilistic Models: Principles and Techniques.* MIT Press, 2006. To appear.

D. Koller and A. Pfeffer. Probabilistic frame-based systems. In *Proceedings of the Fifteenth National Conference on Artificial Intelligence (AAAI)*, pages 580–587, 1998. Available online: `http://citeseer.ist.psu.edu/koller98probabilistic.html`.

S. Kullback and R. A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22(1):79–86, 1951.

Z. Kunda. The case for motivated reasoning. *Psychological Bulletin*, 108(3): 480–498, 1990.

P. M. Lee. *Bayesian Statistics.* Wiley, New York, 1997.

M. R. Lepper and T. R. Shultz. Dissonance. In R. A. Wilson and F. C. Keil, editors, *The MIT Encyclopedia of the Cognitive Sciences.* MIT Press, 2001. Available online: `http://www.psych.mcgill.ca/perpg/fac/shultz/personal/Recent_Publications_files/diss99.pdf`.

M. Lodge and C. Taber. Three steps toward a theory of motivated political reasoning. In A. Lupia, M. McCubbins, and S. Popkin, editors, *Elements of Political Reason: Cognition, Choice, and the Bounds of Rationality*, chapter 9, pages 183–213. 2000.

C. G. Lord, L. Ross, and M. R. Lepper. Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology*, 37(11):2098–2109, 1979.

D. J. C. MacKay. *Information Theory, Inference, and Learning Algorithms.* Cambridge University Press, 2003. Available online: `http://www.inference.phy.cam.ac.uk/mackay/itila/`.

J. L. McClelland. Connectionist models and Bayesian inference. In M. Oaksford and N. Chater, editors, *Rational Models of Cognition.* Oxford University Press, 1998.

A. Miller, H. McHoskey, C. Bane, and T. G. Dowd. The attitude polarization phenomenon: the role of response measure, attitude extremity, and behavioral consequences of reported attitude change. *Journal of Personality and Social Psychology*, 64:561–574, 1993.

S. Mullainathan. A memory-based model of bounded rationality. *Quarterly Journal of Economics*, 2002. Earlier version online: `http://www.economics.harvard.edu/faculty/mullainathan/papers/memory.pdf`.

K. Murphy. A brief introduction to graphical models and Bayesian networks. 1998. Available online: `http://www.cs.ubc.ca/~murphyk/Bayes/bayes.html`.

R. M. Neal and G. E. Hinton. A view of the EM algorithm that justifies incremental, sparse, and other variants. In M. I. Jordan, editor, *Learning in Graphical Models*, pages 355–368. Kluwer Academic Publishers, Dordrecht, 1998. Available online: `http://www.cs.toronto.edu/~radford/em.abstract.html`.

R. S. Nickerson. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, 2(2):175–220, 1998.

R. Nisbett and L. Ross. *Human Inference: Strategies and Shortcomings of Social Judgment*. Prentice Hall, Englewood Cliffs, NJ, 1980.

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

M. Rabin and J. L. Schrag. First impressions matter: A model of confirmatory bias. *Quarterly Journal of Economics*, 114(1):37–82, 1999.

M. Ranney and P. Thagard. Explanatory coherence and belief revision in naive physics. In *Proceedings of the Tenth Annual Conference of the Cognitive Science Society*, pages 426–432. Lawrence Erlbaum, Hillsdale, NJ, 1988.

S. J. Read, E. J. Vanman, and L. C. Miller. Connectionism, parallel constraint satisfaction processes, and Gestalt principles: (Re)Introducing cognitive dynamics to social psychology. *Personality and Social Psychology Review*, 1(1):26–53, 1997.

L. Ross, M. R. Lepper, and M. Hubbard. Perseverance in self perception and social perception: Biased attributional processes in the debriefing paradigm. *Journal of Personality and Social Psychology*, 32:880–892, 1975.

D. E. Rumelhart, P. Smolensky, J. L. McClelland, and G. E. Hinton. Schemata and sequential thought processes in pdp models. In D. E. Rumelhart and J. L. McClelland, editors, *Parallel Distributed Processing: Exploring the Microstructures of Cognition, Volume 2*, chapter 14.

MIT Press, 1986. Available online: `http://www.cnbc.cmu.edu/~jlm/papers/PDP/Chapter14.pdf`.

L. J. Savage. *The Foundations of Statistics.* Wiley, New York, 1954.

R. Shachter. Bayes-Ball: The rational pastime (for determining irrelevance and requisite information in belief networks and influence diagrams). In G. F. Cooper and S. Moral, editors, *Uncertainty in Artificial Intelligence: Proceedings of the Fourteenth Conference*, pages 480–487, San Francisco, 1998. Morgan Kaufmann. Available online: `http://citeseer.ist.psu.edu/shachter98bayesball.html`.

T. R. Shultz and M. R. Lepper. Cognitive dissonance reduction as constraint satisfaction. *Psychological Review*, 103(2):219–240, 1996.

D. Simon, C. J. Snow, and S. J. Read. The redux of cognitive consistency theories: Evidence judgments by constraint satisfaction. *Journal of Personality and Social Psychology*, 86(6):814–837, 2004.

P. Smolensky. Information processing in dynamical systems: Foundations of harmony theory. In D. E. Rumelhart and J. L. McClelland, editors, *Parallel Distributed Processing: Exploring the Microstructures of Cognition, Volume 1.* MIT Press, 1986.

P. Suppes. *Representation and Invariance of Scientific Structures.* CSLI Press, 2001.

P. Thagard. *Coherence in Thought and Action.* MIT Press, 2002.

P. Thagard. Probabilistic networks and explanatory coherence. *Cognitive Science Quarterly*, 1:91–114, 2000. Available online: `http://cogsci.uwaterloo.ca/Articles/Pages/%7FTraditions.html`.

P. C. Wason. Reasoning about a rule. *Quarterly Journal of Experimental Psychology*, 20:273–281, 1968.

P. C. Wason. “On the failure to eliminate hypotheses...”—a second look. In P. N. Johnson-Laird and P. C. Wason, editors, *Thinking: Readings in Cognitive Science*, pages 307–314. Cambridge University Press, 1977.

P. G. Zimbardo, M. Weisenberg, I. Firestone, and B. Levy. Communicator effectiveness in producing public conformity and private attitude change. *Journal of Personality*, 33:233–255, 1965.