# Explainable Structured Bayesian Hypothesis Selection: A Formal Framework for Human-like, Interpretable Belief Formation and Iterative Thought

Authors: Zhang Sulin

## Abstract

We propose a formal, testable framework that models human "selection of viewpoints" as a structured, explainable Bayesian hypothesis-selection process. Each candidate viewpoint is treated as a discrete hypothesis; the cognitive state is a joint selection vector over hypotheses whose probability is governed by evidential potentials and structured relational factors (support, conflict, causation). Rather than casting selection as a linear dimensionality reduction (e.g., PCA), we cast it as a structured MAP/posterior-mass selection problem over combinatorial hypothesis subsets subject to interpretability and parsimony constraints. We present (1) a generative factor-graph formalization, (2) objective functions combining posterior mass, model complexity, and explainability penalties, (3) tractable approximate inference and greedy MAP algorithms suitable for cognitive modeling or AI reasoning modules, and (4) an iterative Candidate–Select–Update (CSU) algorithm that captures the loop of hypothesis generation, adjudication, adoption, and subsequent proposal of novel hypotheses. We outline experimental paradigms (behavioral and computational) to validate the model and discuss applications to explainable AI, chain-of-thought reasoning in LLMs, and algorithmic implementations.

## 1. Introduction

Human reasoning often appears to produce compact sets of interrelated

beliefs (views, hypotheses, or propositions) that jointly explain observations and guide action. These mental "assemblies" are neither single-point beliefs nor raw, unstructured distributions over independent statements; rather they are structured, mutually reinforcing (or conflicting) sets with causal or logical interconnections. A natural formalization is Bayesian: beliefs are probabilistic, updated by evidence. However, classical Bayesian accounts commonly model single hypotheses or factorized distributions and rarely explain how an agent selects an interpretable subset of hypotheses to adopt and reason from.

A tempting but misleading analogy is to treat this selection as principal-component-like dimensionality reduction: choose a small number of high-variance components that span the belief space. We argue that this analogy fails to capture crucial aspects of human cognition: logical relations, causal structure, interpretability, and combinatorial coherency. Instead, we propose a structured Bayesian hypothesis-selection framework in which the cognitive choice is a MAP (or high-mass posterior) selection over combinatorial hypothesis subsets, modulated by semantics-aware factors that enforce interpretability and parsimony.

This paper focuses on the adjudication step—how a candidate hypothesis (or set of hypotheses) is judged for acceptance—treating hypothesis generation as an upstream process that provides proposals possibly influenced by the current adopted set. The framework aims to be mathematically precise, empirically falsifiable, and practically useful for AI systems that need transparent, human-like reasoning.

## 2. Related conceptual frameworks

Our proposal takes inspiration from several literatures: Bayesian models of cognition and "Bayesian brain" hypotheses; graphical models and factor graphs for structured probabilistic reasoning; energy-based models for joint selection; model selection criteria (MAP, MDL, Bayesian model comparison) for parsimony; and recent work on chain-of-thought / iterative reasoning in AI. We position our contribution as an explicit, testable

formalization of subset selection over structured hypothesis spaces with an explicable factor decomposition aimed at interpretability and cognitive plausibility.

## 3. Formal framework

### 3.1 Hypothesis space and selection variables

Let the finite hypothesis space be $H=\{h_1,\ldots,h_n\}$. Each hypothesis is a proposition, rule, or causal conjecture. Define binary selection variables

$Z_i \in \{0,1\}$, $i=1,\ldots,n$.

A cognitive belief state corresponds to a configuration $Z\in\{0,1\}^n$. The interpretation: $Z_i=1$ means the agent "adopts" or acts as though $h_i$ is accepted as part of the working belief set.

### 3.2 Factor-graph generative model

We model the joint probability of selections by a factor graph capturing (i) local evidence for each hypothesis, (ii) pairwise or higher-order relations (supporting/refuting), and (iii) complexity/explainability penalties.

$$P(Z|D) \propto \exp(\Sigma_i \phi_i(Z_i;D) + \Sigma_{i<j} \psi_{ij}(Z_i,Z_j) - \lambda C(Z)),$$

where

$D$ denotes available data/evidence. The unary potentials $\phi_i(1;D)$ measure evidence strength for $h_i$ (log-likelihood ratio or log-odds); $\phi_i(0;D)$ can be set to zero or a baseline.

$\psi_{ij}$ encodes pairwise relations: positive when $h_i$ supports $h_j$, negative when they conflict, structured to reflect causality or logical implication.

$C(Z)$ is a complexity or description-length penalty (e.g., $|Z|_0$ or a weighted

cost reflecting cognitive load or interpretability), and $\lambda \geq 0$ scales parsimony preference.

Higher-order factors (triples or motifs) can be added: $\psi_{ijk}(Z_i, Z_j, Z_k)$ to model triadic coherence.

3.3 Posterior subset selection (MAP / mass criterion)

We define the agent's selected working set as a high-posterior-mass subset. Two natural formal criteria:

MAP selection: $Z^* = \text{argmax}_Z P(Z|D)$.

Posterior mass threshold: Select minimal (or low-complexity) $Z$ such that $P(Z|D) \geq \tau$ for a threshold $\tau$ (e.g., 0.85). This encodes the psychological idea that the agent adopts beliefs only when a coherent bundle exceeds a confidence threshold.

3.4 Interpretability and explainability

Interpretability arises because the potential decomposition exposes local semantic factors: each adopted hypothesis comes with an explainable contribution $\phi_i$ and relational contributions $\psi_{ij}$. We may require additional constraints: each selected $h_i$ must be accompanied by an explanation trace (a short logical/causal chain) whose cost is accounted in $C(Z)$.

3.5 Iterative dynamics: Candidate–Select–Update (CSU)

Agents rarely compute $Z^*$ over the full combinatorial space in one step. Instead we model an iterative loop:

Candidate generation (proposal): propose new hypotheses from a proposal distribution $q(h|Z^{(t)})$, biased by current set.

Adjudication (selection): evaluate $P(Z \cup \{h\}|D)$ approximately and accept if

posterior mass increases beyond threshold or yields higher objective.

Update: if accepted, update working set and use it to propose subsequent hypotheses; if rejected, possibly store hypothesis in a fringe.

This loop is analogous to MCMC or sequential Monte Carlo but oriented to producing an interpretable working set.

## 4. Tractable approximations and algorithmic realizations

Exact MAP over binary vectors is NP-hard in general. We provide practical approximations tailored for cognitive modeling and AI modules.

4.1 Greedy forward selection (GFS)

Start with $Z=\emptyset$. Iteratively add the hypothesis $h$ that maximizes the incremental posterior gain $\Delta(h|Z)=\log P(Z \cup \{h\}|D)-\log P(Z|D)$, stopping when no $h$ yields positive gain or when $P(Z)\geq\tau$. Complexity: $O(nk)$ evaluations for $k$ selections.

4.2 Structured mean-field variational

Relax $Z_i\in\{0,1\}$ to $q_i\in[0,1]$ and minimize $KL(q\|P)$ with structured pairwise terms retained (structured mean-field or structured variational families). This yields interpretable marginals and can be used to extract a ranking of hypotheses by posterior marginals.

4.3 Integer programming / ILP for small sets

For domains with small $n$ or when relations are sparse, formulate MAP as an ILP with linearized pairwise penalties. Leverage off-the-shelf solvers to obtain exact solutions.

4.4 Message passing / loopy belief propagation

When the factor graph is sparse, run belief propagation to approximate marginals; then use heuristics (top-k marginals plus local optimization) to assemble a coherent set.

4.5 Hybrid: proposal + local combinatorial search

Use LLMs or associative memory for proposal generation (candidate pool) and apply a combinatorial optimizer (greedy/ILP/local search) to find high-mass coherent subsets among candidates.

4.6 Pseudo-code: CSU (practical)

Input: hypothesis set H, data D, threshold tau, budget T

Z = empty_set

for t in 1..T:

proposals = propose_candidates(Z, H, D)

evaluate incremental gains Delta(h) for h in proposals

if max Delta(h) > 0:

h_star = argmax Delta(h)

Z = local_optimize(Z U {h_star}) # short local search

if P(Z|D) >= tau: break

else:

break

return Z

## 5. Experimental validation

We propose a two-pronged validation program: (A) Behavioral experiments with human subjects to test cognitive predictions, (B) Computational experiments for AI implementations.

### 5.1 Behavioral predictions and tasks

Confidence thresholding: Present subjects with bundles of interrelated claims and evidence; predict that acceptance of a bundle depends on joint coherence rather than individual statement strength. Measure minimal bundle size and joint-support needed to cross subjective confidence thresholds.

Proposal dependence: Given a fixed evidence set, vary the sequence of candidate proposals; predict path dependence in adopted belief sets (due to iterative CSU dynamics).

Explanatory parsimony: Subjects will prefer bundles with lower description cost (fewer explanatory primitives) even if marginal likelihoods are similar.

Collect behavioral metrics: subjective confidence, choice of accepted statements, reaction times, free explanations (to assess interpretability traces).

### 5.2 Computational experiments

Synthetic factor-graph worlds: generate ground-truth hypothesis interrelations, sample evidence, and compare algorithms (GFS, variational, ILP, baseline PCA-like ranking) on recovery of ground-truth adopted sets.

LLM-based agents: implement proposal module with LLM prompts; use combinatorial adjudication to form working sets; evaluate on reasoning

benchmarks requiring multi-fact coherence.

Human-in-the-loop evaluation: compare AI explanations extracted from adopted sets against human explanations on benchmark tasks (e.g., abductive reasoning datasets).

Metrics: precision/recall for recovered adopted hypotheses, coherence score (aggregate pairwise support), explanation length and human-rated interpretability.

## 6. Applications to AI

Explainable chain-of-thought: rather than unstructured token chains, produce compact adopted hypothesis sets with attached causal/logical traces usable for transparent reasoning and verification.

Belief-state maintenance in agents: maintain an explicit, low-complexity set of state hypotheses for planning and learning.

Alignment and safety: explicit hypothesis sets facilitate human inspection, constraint injection, and safer behavior.

## 7. Discussion and limitations

The factorial model trades expressivity for computational complexity. Real cognition likely uses heuristics (e.g., focused proposal, associative search) that our CSU algorithm models qualitatively.

Mapping natural-language beliefs to a finite hypothesis space requires strong representational assumptions (symbolic parsing or semantic parsers). Hybrid neuro-symbolic front-ends are a promising direction.

The framework assumes evidence can be scored per hypothesis; in practice evidence may be ambiguous and shared among hypotheses. This is

addressed by relational factors but empirical calibration is needed.

## 8. Conclusion

We presented a structured Bayesian hypothesis-selection framework that models human adoption of interpretable belief subsets via posterior-mass-based selection with explicit relational and complexity factors. The framework clarifies why PCA is a misleading analogy and provides algorithmic realizations (CSU and approximations) applicable to cognitive modeling and AI systems demanding transparent reasoning. The next steps are empirical validation and implementation in LLM-guided reasoning modules.

## Suggested further work

Formal analysis of identifiability: when is the true adopted subset recoverable from observations?

Richer proposal models: generative neural proposals conditioned on current set.

Integration with memory and attention constraints: model finite cognitive resource allocation.

Case studies: law, medical diagnosis, and scientific reasoning.

## Acknowledgments

Preprint prepared interactively with GPT-5 Thinking mini.

## References


Amit, D. J. (1989). Modeling Brain Function: The World of Attractor Neural Networks. Cambridge University Press.

Barber, D. (2012). Bayesian Reasoning and Machine Learning. Cambridge University Press.

Bishop, C. M. (2006). Pattern Recognition and Machine Learning. Springer.

Burnham, K. P., & Anderson, D. R. (2002). Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach (2nd ed.). Springer.

Doya, K., Ishii, S., Pouget, A., & Rao, R. P. N. (Eds.). (2007). Bayesian Brain: Probabilistic Approaches to Neural Coding. The MIT Press.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. The MIT Press.

Grünwald, P. D. (2007). The Minimum Description Length Principle. The MIT Press.

Koller, D., & Friedman, N. (2009). Probabilistic Graphical Models: Principles and Techniques. The MIT Press.

Mackay, D. J. C. (2003). Information Theory, Inference, and Learning Algorithms. Cambridge University Press.

Murphy, K. P. (2012). Machine Learning: A Probabilistic Perspective. The MIT Press.

Oaksford, M., & Chater, N. (2007). Bayesian Rationality: The Probabilistic Approach to Human Reasoning. Oxford University Press.

Pearl, J. (1988). Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann.

Pearl, J. (2000). Causality: Models, Reasoning, and Inference. Cambridge University Press.

Pouget, A., Beck, J. M., Ma, W. J., & Latham, P. E. (2013). Probabilistic Brains: Knowns and Unknowns. The MIT Press.

Russell, S., & Norvig, P. (2020). Artificial Intelligence: A Modern Approach (4th ed.). Pearson.

Tenenbaum, J. B., Griffiths, T. L., & Kemp, C. (2006). Theory-Based Bayesian Models of Inductive Learning and Reasoning. Trends in Cognitive Sciences.

Dayan, P., & Abbott, L. F. (2001). Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems. MIT Press.

Frey, B. J., & Jojic, N. (2005). A comparison of algorithms for inference and learning in probabilistic graphical models. IEEE Transactions on Pattern Analysis and Machine Intelligence, 27(9), 1392–1416.

Kschischang, F. R., Frey, B. J., & Loeliger, H.-A. (2001). Factor graphs and the sum-product algorithm. IEEE Transactions on Information Theory, 47(2), 498–519.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., & Zhou, D. (2022). Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. Advances in Neural Information Processing Systems, 35.