

Predictive Coding for Hierarchical Belief Systems

For a discrete, event-sourced belief system, predictive coding is best understood as an **inference schedule over a generative model**, not as a replacement for probabilistic modeling itself. In its classical form, it is a layered Gaussian/Laplace approximation that passes **precision-weighted prediction errors** upward and predictions downward. That is materially different from factor-graph belief propagation, which passes **functions over variables** and, on trees, computes exact marginals. For continuous linear-Gaussian hierarchies, the two can agree at the posterior mean or mode. For discrete events, symbolic structure, irregular arrivals, or dynamic graph growth, the classical Rao–Ballard form does not transfer directly; the literature increasingly moves either to arbitrary-distribution predictive coding, arbitrary graph topologies, or to factor-graph variational message passing for the discrete parts and predictive-coding-like updates for the continuous parts. ¹

Formal comparison

Consider a static two-level hierarchy with observation x , lower latent z_1 , and upper latent z_2 . A standard predictive-coding generative model is

$$\begin{aligned} p(z_2) &= \mathcal{N}(z_2; \mu_2^{\text{prior}}, \Sigma_2^{\text{prior}}) \\ p(z_1 | z_2) &= \mathcal{N}(z_1; f_2(z_2), \Sigma_2) \\ p(x | z_1) &= \mathcal{N}(x; f_1(z_1), \Sigma_1). \end{aligned}$$

Under the Laplace or delta-posterior approximation used in classical predictive coding, $q(z_1, z_2) \approx \delta(z_1 - \mu_1) \delta(z_2 - \mu_2)$, inference minimizes a variational free-energy surrogate equivalent, up to constants, to the negative joint log density evaluated at the current means:

$$F(\mu_1, \mu_2) = \frac{1}{2}(x - f_1(\mu_1))^\top \Pi_1(x - f_1(\mu_1)) + \frac{1}{2}(\mu_1 - f_2(\mu_2))^\top \Pi_2(\mu_1 - f_2(\mu_2)) + \frac{1}{2}(\mu_2 - \mu_2^{\text{prior}})^\top \Pi_3(\mu_2 - \mu_2^{\text{prior}}) + C$$

with precisions $\Pi_i = \Sigma_i^{-1}$. Define local prediction errors

$$\varepsilon_0 = \Pi_1(x - f_1(\mu_1)), \quad \varepsilon_1 = \Pi_2(\mu_1 - f_2(\mu_2)), \quad \varepsilon_2 = \Pi_3(\mu_2 - \mu_2^{\text{prior}}).$$

Gradient-flow updates are then

$$\dot{\mu}_1 = J_{f_1}(\mu_1)^\top \varepsilon_0 - \varepsilon_1, \quad \dot{\mu}_2 = J_{f_2}(\mu_2)^\top \varepsilon_1 - \varepsilon_2,$$

where J_{f_i} is the Jacobian of f_i . In the linear case, $f_1(z_1) = W_1 z_1$ and $f_2(z_2) = W_2 z_2$, this becomes

$$\begin{aligned} \dot{\mu}_1 &= W_1^\top \Pi_1(x - W_1 \mu_1) - \Pi_2(\mu_1 - W_2 \mu_2), \\ \dot{\mu}_2 &= W_2^\top \Pi_2(\mu_1 - W_2 \mu_2) - \Pi_3(\mu_2 - \mu_2^{\text{prior}}). \end{aligned}$$

The distinctive point is architectural: the upward-traveling quantity is the precision-weighted residual ε_i , not the raw state or raw observation. ²

The same hierarchy written for exact Bayesian inference on a factor graph is

$$p(x, z_1, z_2) = p(z_2) p(z_1 | z_2) p(x | z_1).$$

On a tree, sum-product belief propagation passes messages that are **functions of the variable values**, not signed residuals:

$$\begin{aligned} m_{z_1 \leftarrow x}(z_1) &= p(x_{\text{obs}} | z_1), \\ m_{z_2 \leftarrow z_1}(z_2) &= \int p(z_1 | z_2) m_{z_1 \leftarrow x}(z_1) dz_1, \\ b(z_2) &\propto p(z_2) m_{z_2 \leftarrow z_1}(z_2), \\ b(z_1) &\propto p(x_{\text{obs}} | z_1) \int p(z_1 | z_2) p(z_2) dz_2. \end{aligned}$$

In general factor-graph form, variable-to-factor messages are products of incoming messages from neighboring factors, and factor-to-variable messages are sums or integrals of the local factor times incoming cavity messages from the other neighboring variables. On trees this is exact; on loopy graphs it is approximate. ³

The formal differences are therefore direct:

- **Object being optimized:** predictive coding performs gradient descent on a variational free-energy surrogate over latent means; sum-product BP computes exact marginals on trees and fixed points of loopy BP correspond to stationary points of the Bethe free energy, not the same objective. ⁴
- **Message content:** predictive coding upward messages are local precision-weighted residuals ε_i ; BP messages are functions or distributions $m(\cdot)$. ⁵
- **Posterior representation:** predictive coding in its classical form tracks means, and only implicitly covariance through precision parameters; BP represents beliefs over states and, in Gaussian BP, propagates information sufficient for marginals and covariances. ⁵
- **Assumptions:** predictive coding is usually derived for differentiable hierarchical generative maps with Gaussian noise and iterative relaxation to equilibrium; BP only requires factorization and supports discrete, continuous, or mixed factors. ⁶

A technical correction is important. “Standard bottom-up hierarchical Bayes” is not, in its exact form, purely bottom-up. Exact inference on a hierarchy already uses both upward likelihood messages and downward prior or cavity messages. The real contrast is therefore **not** bidirectionality versus unidirectionality. The contrast is **residual passing with local gradient descent** versus **distribution passing with sum-product integration**. ⁷

Precision weighting

In predictive coding, the error signal from level i is not just $e_i = v_i - \hat{v}_i$. It is the **precision-weighted** error

$$\varepsilon_i = \Pi_i(v_i - \hat{v}_i),$$

where $\Pi_i = \Sigma_i^{-1}$ is the precision of the generative conditional for that level, $p(v_i | v_{i+1})$. At the sensory level this is the likelihood noise of $p(x | z_1)$. At higher levels it is the uncertainty of the inter-level mapping $p(z_1 | z_2)$. This is why predictive coding can down-weight a large lower-level mismatch if the model says that mapping is noisy, while strongly amplifying a smaller mismatch from a highly reliable mapping. ⁸

A concrete scalar example shows the effect. Suppose two lower-level channels both predict the same parent variable with unit Jacobian:

- channel A: raw error $e_A = 10$, variance $\Sigma_A = 100$, precision $\Pi_A = 0.01$
- channel B: raw error $e_B = 1$, variance $\Sigma_B = 0.25$, precision $\Pi_B = 4$.

Then the predictive-coding messages are

$$\varepsilon_A = 0.01 \times 10 = 0.1, \quad \varepsilon_B = 4 \times 1 = 4.$$

So the **low-precision, high-magnitude** mismatch is almost ignored, while the **high-precision, low-magnitude** mismatch dominates the parent update. This is exactly the behavior classical predictive-coding papers attribute to variance normalization: noisier channels exert less influence even when their raw residual is large. ⁹

If one performs the exact Gaussian Bayesian update with the same model assumptions, the same weighting appears in the final posterior. With a zero-mean unit-precision prior on the parent variable z ,

$$p(z) = \mathcal{N}(0, 1), \quad x_A = 10, \Sigma_A = 100, \quad x_B = 1, \Sigma_B = 0.25,$$

the posterior precision is

$$1 + 0.01 + 4 = 5.01,$$

and the posterior mean is

$$\mathbb{E}[z | x_A, x_B] = \frac{0 \cdot 1 + 10 \cdot 0.01 + 1 \cdot 4}{5.01} = \frac{4.1}{5.01} \approx 0.818.$$

So exact Bayesian inference also favors the smaller but more precise signal. The key distinction is therefore not that predictive coding violates Bayes or obtains a different optimum. The distinction is that predictive coding performs the reliability weighting **locally and explicitly in the transmitted residual**. A naive bottom-up architecture that passed raw magnitudes upward before reliability correction would react to 10 more strongly than to 1; predictive coding does not. ¹⁰

Design-wise, this means a hierarchical belief system built in predictive-coding style must store precision parameters on **every inter-level conditional**, not just at the sensor boundary. In a discrete event system, that corresponds to confidence on event-to-state, state-to-aggregate, and aggregate-to-hypothesis links separately. ¹¹

Discrete and structured domains

The machine-learning literature does support predictive coding beyond early visual models, but the strongest results are about **differentiable networks and arbitrary graph topologies**, not about symbolic knowledge-graph reasoning in the strict ontology sense. Whittington and Bogacz showed that a predictive-coding network with local Hebbian updates can perform supervised learning and that, for certain parameter settings, its weight updates converge to those of backpropagation. That result is about credit assignment in layered differentiable networks, not directly about symbolic inference. ¹²

The later theory paper coauthored by Tommaso Salvatori shows a broader point: for standard predictive-coding networks trained with “prospective configuration,” the inference equilibrium can be interpreted as interpolating between feedforward values and target-propagation-style local targets, with the balance controlled by feedforward and feedback precisions; learning can be interpreted as a form of generalized expectation-maximization, and the authors prove convergence to critical points of the backprop loss. This is an important result if the question is whether predictive coding is a serious learning architecture rather than only a neuroscientific metaphor. ¹³

Salvatori’s NeurIPS 2022 “PC graphs” paper is the strongest evidence that predictive-coding message passing is not inherently tied to a fixed layer stack. It defines predictive-coding graphs on **any directed graph**, states that the same graph can be queried in multiple ways by clamping different nodes, and treats different queries as conditional expectations over the graph’s variables. That is directly relevant to belief systems whose structure is not a simple chain. However, the state variables in that work remain continuous-valued nodes optimized by minimizing local squared prediction errors; it is a structured differentiable architecture, not a symbolic theorem prover or ontology reasoner. ¹⁴

The most relevant extension for discrete events is the work on predictive coding beyond Gaussian distributions. That line starts from the observation that classical predictive coding is a hierarchical Gaussian generative model and generalizes it to arbitrary probability distributions, explicitly to support architectures such as transformers and conditional language models. This matters because once observations are categorical, Bernoulli, count-valued, or otherwise non-Gaussian, the “prediction error” is no longer naturally a Euclidean subtraction. The operative local signal becomes a gradient of a log-likelihood or negative log-probability term, not a raw signed residual. ¹⁵

There is also a newer line on structured world models via causal graphs. The 2024 work on predictive coding beyond correlations shows that predictive-coding-style inference can be modified to compute interventions and counterfactuals on causal graphs, including cases where graph structure must be inferred from data. This is the closest retrieved literature to “structured world representations” in the sense of explicit relation structure. It is still closer to causal or Bayesian networks than to ontology reasoning or knowledge-graph completion. ¹⁶

Within the literature retrieved here, I did **not** find an established line that uses classical Rao–Ballard predictive-coding message passing as the main inference engine for knowledge graphs or ontologies in the way factor-graph methods are used for probabilistic relational models. The closest retrieved lines are: arbitrary directed PC graphs; predictive coding on graph-representation-learning tasks; and causal-graph intervention or counterfactual inference. That is evidence that the framework can be generalized structurally, but not that the symbolic or ontology case is already solved. ¹⁷

The continuous-observation formulation breaks in several specific ways when observations are discrete events arriving irregularly. First, the local error term $x - \hat{x}$ ceases to be the right statistic for categorical or event-count data; one needs likelihood-specific gradients or message updates instead. Second, a classical predictive-coding network usually assumes either static relaxation to a fixed point or continuous-time generalized coordinates with local sampling in time bins; irregular event arrivals violate that clean separation between “infer until convergence” and “receive the next datum.” Third, absence of events can itself be informative in event streams, which pushes the model toward hazard, survival, or point-process likelihoods rather than additive Gaussian noise. Fourth, many symbolic constraints are combinatorial and exact, whereas quadratic error energies are soft and continuous. The mixed discrete/continuous active-inference literature addresses this by using factor-graph message passing for discrete state spaces and predictive-coding-like gradient dynamics for continuous states, on the same overall generative model. ¹⁸

Relationship to loopy belief propagation

Loopy belief propagation and predictive coding are related at a high level but are not generally equivalent. The cleanest statement is:

- loopy BP performs approximate marginal inference on a factor graph, and its fixed points are stationary points of the Bethe free energy;
- predictive coding performs gradient descent on a variational free-energy objective over latent means under a particular generative parameterization and approximation family;
- both are local message-passing schemes for generative models, but they optimize different free-energy functionals and represent different posterior objects. ¹⁹

They do coincide in special cases. For a **tree-structured linear-Gaussian** model, the predictive-coding objective is quadratic, the Laplace approximation is exact, and the converged means can equal the exact posterior mean, which is also the posterior mode. In that narrow case, predictive coding and Gaussian BP agree on the point estimate. However, Gaussian BP still carries covariance information and produces marginals, while vanilla predictive coding generally does not unless extended with second-order structure. Outside the linear-Gaussian case, especially for discrete or multimodal posteriors, the equivalence breaks. ²⁰

The most informative bridge is the active-inference or factor-graph literature associated with Karl Friston ²¹. There, deep generative models are written as factor graphs that can contain both discrete and continuous states. The discrete-state parts are updated by belief propagation or variational message passing, while the continuous-state parts admit predictive-coding-like gradient flows. This is an argument for viewing predictive coding and BP as **different update schemes within a broader variational-inference family**, rather than as two names for the same algorithm. ²²

Convergence differs sharply. On a tree, BP is exact and converges after a finite message schedule. On loopy graphs, there is no general unconditional convergence guarantee. A standard sufficient condition for binary pairwise models is that the spectral radius of an interaction matrix be strictly less than 1; more generally, the literature frames convergence through contraction conditions and uniqueness of fixed points. Yedidia, Freeman, and Weiss also show that on loopy graphs BP fixed points correspond to stationary points of the Bethe approximation, which explains both its successes and its failures. ²³

For a dynamic belief graph whose structure changes online and whose number of levels is not fixed at design time, those theorems do not carry over directly, because they assume a **fixed graph and fixed factors during iteration**. The most defensible engineering translation is therefore an inference: if structure can change while messages are still relaxing, neither classical loopy-BP convergence results nor classical predictive-coding fixed-point arguments apply without modification. The practical implication is to freeze topology during an inference epoch, use damping or asynchronous scheduling, and then restart or locally re-solve around graph edits. The predictive-coding-graph literature supports arbitrary fixed topologies; it does not, by itself, establish convergence for online graph growth. ²⁴

Failure modes and safeguards

The central failure mode is the one the psychosis and hallucination literature makes explicit: if high-level priors become too precise, or if bottom-up evidence is assigned too little precision, contradictory sensory evidence is “explained away” rather than allowed to move the posterior. The result is a hallucination-like regime in which top-down prediction suppresses legitimate lower-level evidence. Reviews of hallucinations and psychosis frame this either as overly strong high-level priors, imbalanced precision assignment across levels, or a failure of sensory attenuation with compensatory top-down precision increases; they also stress that the pathology is hierarchical and that strong-prior versus weak-prior findings can coexist at different levels of the system. ²⁵

The first concrete safeguard is **adaptive precision control of prediction-error units**, often interpreted as attention. In this account, the system does not merely compute prediction errors; it separately estimates their precision and modulates the gain on those units so that reliable bottom-up channels remain influential. That is the formal mechanism by which predictive coding avoids letting a confident but unreliable top-down model silence trustworthy input. This is also the main route by which attention enters the predictive-coding literature. ²⁶

The second concrete safeguard is **sensory attenuation or corollary-discharge gating for self-generated input**. The proposal is not to silence all bottom-up evidence, but specifically to attenuate the precision of the expected sensory consequences of one’s own actions while preserving sensitivity to externally caused signals. Failures of this mechanism have been proposed for agency disturbances and hallucinations, and empirical work has linked deficient predictive coding to hallucination-related hyperactivity in sensory cortex. In engineering terms, the analogous mechanism is to mark self-caused, model-predicted event channels and lower their upward precision selectively, rather than globally. ²⁷

The third concrete safeguard is **epistemic action or active sampling**. In the active-inference literature, policies are scored partly by expected information gain, so the system seeks observations that reduce ambiguity instead of persisting in a prior-driven interpretation. This matters for hallucination prevention because when the evidence is weak or conflicting, the model’s remedy is not simply to increase prior confidence; it is to gather more informative data. For an event-sourced system, the analogue is to trigger disambiguating queries, targeted sensing, or dependency checks when high-level and low-level messages disagree under high uncertainty. ²⁸

Design implications

The literature supports a specific conclusion for an event-sourced hierarchical belief system.

A **pure classical predictive-coding stack is not, by itself, the most general architecture** for this domain. It is strongest when the hierarchy is differentiable, approximately Gaussian, and can relax iteratively between observations. It is weaker when observations are discrete symbols or sparse irregular events, when exact marginals matter, when graph structure changes online, or when the system needs explicit logical or relational constraints. In those cases, factor graphs and variational or belief-propagation methods remain the cleaner formalism. Predictive coding then becomes a useful **local solver** or **continuous-state subroutine**, not the whole architecture. ²⁹

The design changes that follow are substantial.

Represent the system as an explicit generative graph over events, latent states, aggregates, and times. For continuous submodels, let parent nodes send predictions downward and let child nodes return precision-weighted residuals upward, as in predictive coding. For discrete event factors, do not force a subtraction-based residual. Use a categorical, Bernoulli, count, or hazard likelihood and pass either variational messages or the local gradient of the appropriate negative log-likelihood. This follows the direction taken by arbitrary-distribution predictive coding and by mixed discrete/continuous active-inference models. ³⁰

Attach a reliability parameter to **every** conditional, not only to raw sensors. In a discrete event system this means maintaining uncertainty on event-detection quality, on event-to-state mappings, on state-to-aggregate mappings, and on cross-level temporal predictions. That is the correct translation of precision weighting from the continuous literature into a hierarchical belief system. ⁸

Do not rely on a fixed depth or a globally synchronized relaxation schedule. If the graph can grow as new entities or abstractions are created, process inference in epochs with frozen structure, use damping or asynchronous local updates, and treat topology edits as objective changes that require re-initialization or local re-solving. This is not a theorem from the retrieved literature; it is the strongest design inference consistent with the convergence results and their assumptions. ³¹

Finally, if the system must avoid hallucination-like suppression of legitimate events, add explicit mechanisms analogous to the literature's safeguards: precision floors on trusted bottom-up channels, selective attenuation only for self-predicted or low-reliability channels, and an escalation policy that requests more evidence when precision-weighted conflicts persist rather than merely increasing prior strength. Those choices are engineering analogues of attention, sensory attenuation, and epistemic action. ³²

Open questions and limitations

The literature retrieved here is strong on hierarchical Gaussian predictive coding, on its relation to backpropagation, on arbitrary graph topologies, and on mixed discrete/continuous active-inference factor graphs. It is materially thinner on predictive-coding message passing for **knowledge graphs, ontologies, or symbolic world models** as first-class reasoning systems. The closest retrieved work broadens predictive coding to arbitrary directed graphs and causal graphs, not to ontology reasoning in the semantic-web or probabilistic-relational sense. ³³

That gap matters for architecture choice. If the belief system needs exact symbolic consistency, discrete relational constraints, calibrated marginals, and dynamic graph surgery, the current literature favors a

hybrid: factor-graph or variational message passing for the discrete structure, predictive-coding-style local residual dynamics for continuous embeddings or continuous latent submodels. The literature supports that hybrid most directly; it does not yet support replacing the full structured inference stack with classical Rao-Ballard predictive coding alone. ³⁴

¹ ² ⁵ ⁶ ⁸ ⁹ ¹¹ https://www.tnu.ethz.ch/fileadmin/user_upload/teaching/cpcourse/2020/Literature/Bogacz_2017.pdf

https://www.tnu.ethz.ch/fileadmin/user_upload/teaching/cpcourse/2020/Literature/Bogacz_2017.pdf

³ ⁴ ⁷ ¹⁰ ¹⁹ <https://www.cs.princeton.edu/courses/archive/spr06/cos598C/papers/YedidaFreemanWeiss2004.pdf>

<https://www.cs.princeton.edu/courses/archive/spr06/cos598C/papers/YedidaFreemanWeiss2004.pdf>

¹² <https://www.bndu.ox.ac.uk/papers/approximation-error-backpropagation-algorithm-predictive-coding-network-local-hebbian>

<https://www.bndu.ox.ac.uk/papers/approximation-error-backpropagation-algorithm-predictive-coding-network-local-hebbian>

¹³ <https://arxiv.org/abs/2207.12316>

<https://arxiv.org/abs/2207.12316>

¹⁴ ¹⁷ ³³ https://proceedings.neurips.cc/paper_files/paper/2022/file/f9f54762cbb4fe4dbffdd4f792c31221-Paper-Conference.pdf

https://proceedings.neurips.cc/paper_files/paper/2022/file/f9f54762cbb4fe4dbffdd4f792c31221-Paper-Conference.pdf

¹⁵ ³⁰ <https://arxiv.org/abs/2211.03481>

<https://arxiv.org/abs/2211.03481>

¹⁶ <https://proceedings.mlr.press/v235/salvatori24a.html>

<https://proceedings.mlr.press/v235/salvatori24a.html>

¹⁸ ²¹ ²² ²⁹ ³⁴ https://discovery.ucl.ac.uk/id/eprint/1574278/7/Friston%20VoR%20netn_a_00018.pdf

https://discovery.ucl.ac.uk/id/eprint/1574278/7/Friston%20VoR%20netn_a_00018.pdf

²⁰ <https://journals.plos.org/ploscompbiol/article/file?id=10.1371%2Fjournal.pcbi.1000211&type=printable>

<https://journals.plos.org/ploscompbiol/article/file?id=10.1371%2Fjournal.pcbi.1000211&type=printable>

²³ ²⁴ ³¹ https://staff.fnwi.uva.nl/j.m.mooij/articles/uai2005_1139.pdf

https://staff.fnwi.uva.nl/j.m.mooij/articles/uai2005_1139.pdf

²⁵ <https://www.sciencedirect.com/science/article/pii/S1364661318302821>

<https://www.sciencedirect.com/science/article/pii/S1364661318302821>

²⁶ ³² https://www.fil.ion.ucl.ac.uk/spm/doc/papers/Attention_uncertainty_and_free-energy.pdf

https://www.fil.ion.ucl.ac.uk/spm/doc/papers/Attention_uncertainty_and_free-energy.pdf

²⁷ <https://www.frontiersin.org/journals/psychiatry/articles/10.3389/fpsy.2013.00047/full>

<https://www.frontiersin.org/journals/psychiatry/articles/10.3389/fpsy.2013.00047/full>

²⁸ <https://chrismathys.com/wp-content/uploads/2015/05/Friston-et-al.-2015-Active-inference-and-epistemic-value.pdf>

<https://chrismathys.com/wp-content/uploads/2015/05/Friston-et-al.-2015-Active-inference-and-epistemic-value.pdf>