From Collective Predictive Coding to Human–AI Synergy: A Path for Symbiotic Coordination

Yusuke Hayashi [1]

Abstract

Collective Predictive Coding (CPC) broadens the classical framework of Predictive Coding (PC) by positing a shared external representation (e.g., language, symbols, or common knowledge) that couples agents in a multi-agent setting. Prior work has shown that both PC and CPC can be analyzed from a Bayesian perspective, with their updates expressible via Langevin equations under suitable assumptions. Notably, the CPC-derived Langevin dynamics introduce an additional potential term that can be viewed as an "external force," capturing how shared symbols steer the collective.

However, to fully grasp why substituting Bayesian updates with Langevin dynamics is valid, one must recognize that the corresponding Fokker–Planck equation converges to the same posterior distribution implied by Bayesian inference. In this paper, we restore and expand the technical details linking Bayesian updating, Fokker–Planck convergence, and the emergence of the CPC-specific force term. We also offer a more thorough discussion of how each free-energy component in PC and CPC is derived, why it matters for multi-agent coordination, and what limitations arise from communication constraints and symbol emergence. These clarifications provide a stronger foundation for leveraging CPC to orchestrate hybrid human–AI collectives via shared external media.

1. Introduction

In multi-agent scenarios, especially those involving humans and AI, collective dynamics often exhibit complexities that cannot be reduced to independent agent behaviors. Traditional *Predictive Coding* (PC) accounts for each agent (or system) updating its internal model by minimizing prediction errors (Rao & Ballard, 1999; Friston & Kiebel, 2009). Yet, humans and AI systems frequently share external structures—such as languages, symbols, or knowledge repositories—that act as coupling mechanisms.

Collective Predictive Coding (CPC) (Taniguchi, 2024) integrates these shared elements by adding a collective regularization term to the free energy, thereby encouraging agents to align through a common representation w. Importantly, in a Bayesian viewpoint, the CPC update equations can be mapped to stochastic differential equations (Langevin dynamics) with an extra potential term functioning like a soft "external force" on agents.

Despite the intrigue of this result, questions remain:

- **Free-Energy Components:** What exactly are the terms in PC vs. CPC, and how do they differ?
- **Bayesian-to-Langevin Equivalence:** Why is it valid to use Langevin dynamics to discuss convergence under Bayesian updates?
- **Fokker–Planck Equations:** How do these equations show that the equilibrium distribution is indeed the posterior distribution under Bayesian inference, ensuring that the Langevin perspective is consistent?
- **Practical Complexity:** How do real-world constraints such as imperfect communication or emergent symbols affect the CPC force term?

In what follows, we revisit the derivation of PC and CPC free energy, show how the corresponding Bayesian updates lead to a Fokker–Planck description whose equilibrium is the target posterior, and illustrate how this ties neatly into the Langevin formulation. By explicitly analyzing the role of the Fokker–Planck equation, we clarify why replacing Bayesian updates with Langevin dynamics is justified. We then discuss the implications of CPC's extra coupling term and highlight potential application areas for hybrid human–AI systems, as well as research gaps that need addressing.

[1] AI Alignment Network (ALIGN), Tokyo, Japan. Correspondence to: Yusuke Hayashi <hayashi@aialign.net>.

Proceedings of the 2^{nd} Tokyo AI Safety Conference, Tokyo, Japan.

2. Predictive Coding and Collective Predictive Coding

In this section, we reintroduce the detailed derivations of Predictive Coding (PC) and its extension, Collective Predictive Coding (CPC). The aim is to show how each term in the free energy arises, why it connects to Bayesian inference, and what changes in the multi-agent setting that includes a shared representation w.

2.1. Predictive Coding (PC)

Generative Model. Predictive Coding traditionally assumes a probabilistic generative model where latent states z^k generate observations o^k for an agent k. Let:

$$p_\theta(\mathbf{z}, \mathbf{o} \mid \mathbf{a}, \mathbf{C}) \;=\; \prod_k p_\theta(z^k \mid a^k)\, p_\theta(o^k \mid z^k, a^k, C^k).$$

- $p_\theta(z^k \mid a^k)$ specifies how the agent's latent state depends on actions or control variables.
- $p_\theta(o^k \mid z^k, a^k, C^k)$ describes how observations arise from latent states and rewards/context C^k.

Inference Model and Free Energy. Agents maintain a variational distribution $q_\phi(\mathbf{z}, \mathbf{o} \mid \mathbf{a}, \mathbf{C})$ to approximate the true posterior. The PC free energy can be written as the Kullback–Leibler divergence between the generative model and the variational distribution:

$$F_{\mathrm{PC}}(\theta, \phi) \;=\; D_{\mathrm{KL}}\Big[q_\phi(\mathbf{z}, \mathbf{o} \mid \mathbf{a}, \mathbf{C}) \,\Big\|\, p_\theta(\mathbf{z}, \mathbf{o} \mid \mathbf{a}, \mathbf{C})\Big].$$

Expanding, we obtain two main sums:

$$F_{\mathrm{PC}}(\theta, \phi) = \sum_k \underbrace{\mathbb{E}_q\Big[\ln \frac{q_\phi(z^k \mid o^k, a^k)}{p_\theta(z^k \mid a^k)}\Big]}_{\text{(A) Individual regularization}} + \sum_k \underbrace{\mathbb{E}_q\Big[\ln \frac{q_\phi(o^k \mid C^k)}{p_\theta(o^k \mid z^k, a^k, C^k)}\Big]}_{\text{(B) Prediction accuracy (surprise)}}. \tag{1}$$

1. **(A) Individual Regularization:** This term represents the KL divergence between the posterior $q_\phi(z^k \mid o^k, a^k)$ and the prior $p_\theta(z^k \mid a^k)$. Minimizing it reduces the complexity of the latent states, preventing them from straying too far from prior expectations.
2. **(B) Prediction Accuracy (Surprise):** This term captures how well the model predicts observations. $\mathbb{E}_q[\ln p_\theta(o^k \mid z^k, a^k, C^k)]$ is akin to a log-likelihood, and the difference with $\ln q_\phi(o^k \mid C^k)$ measures how surprising the observations are under the model.

2.2. Collective Predictive Coding (CPC)

Shared Representation w. When N agents share a symbolic or language-based representation w, the generative model becomes:

$$p_\theta(w, \mathbf{z}, \mathbf{o} \mid \mathbf{a}, \mathbf{C}) \;=\; p_\theta(w) \prod_k p_\theta\big(z^k \mid w, a^k\big)\, p_\theta\big(o^k \mid z^k, a^k, C^k\big).$$

The inference model analogously introduces $q_\phi(w \mid \mathbf{z})$, ensuring that w is inferred from the states of all agents.

CPC Free Energy. The CPC free energy is:

$$F_{\mathrm{CPC}}(\theta, \phi) = D_{\mathrm{KL}}\Big[q_\phi(w, \mathbf{z}, \mathbf{o} \mid \mathbf{a}, \mathbf{C}) \,\|\, p_\theta(w, \mathbf{z}, \mathbf{o} \mid \mathbf{a}, \mathbf{C})\Big].$$

Its expanded form highlights an additional *collective regularization* term:

$$F_{\mathrm{CPC}}(\theta, \phi) = \underbrace{\mathbb{E}_q\Big[\ln \frac{q_\phi(w \mid \mathbf{z})}{p_\theta(w)}\Big]}_{\text{(C) Collective term}} + \sum_k \underbrace{\mathbb{E}_q\Big[\ln \frac{q_\phi(z^k \mid w, o^k, a^k)}{p_\theta(z^k \mid w, a^k)}\Big]}_{\text{(D) Indiv. reg. w.r.t. } w} + \sum_k \underbrace{\mathbb{E}_q\Big[\ln \frac{q_\phi(o^k \mid C^k)}{p_\theta(o^k \mid z^k, a^k, C^k)}\Big]}_{\text{(E) Prediction accuracy (surprise)}}. \tag{2}$$

- **(C) Collective Regularization:** This new term reflects how the posterior $q_\phi(w \mid \mathbf{z})$ might deviate from the prior $p_\theta(w)$. Essentially, it measures the complexity or "collective cost" of adjusting the external representation w to align with the agents' latent states $\mathbf{z}$.
- **(D) Individual Regularization (with w):** Each agent's latent state must be consistent not only with local data (o^k, a^k) but also with the shared w.
- **(E) Prediction Accuracy:** Remains similar to standard PC but is now influenced indirectly by w.

2.3. Why Does CPC Introduce a New Coupling?

In standard PC, each agent's updates are driven purely by its own observations and priors. By contrast, CPC agents also track how consistent w is with their states and others', effectively adding a "social" or "collective" potential to the free-energy surface. As we will see, this extra term reappears in the Langevin equations as an external force.

3. Bayesian Updating, Langevin Equations, and the Fokker–Planck Perspective

This section clarifies why Bayesian updating—as implied by free-energy minimization—can be cast as a Langevin process, and how the Fokker–Planck equation formalizes the convergence of this process to the posterior distribution. By restoring these details, we connect the free-energy perspective more explicitly to the idea that the final stationary distribution is indeed the Bayesian posterior.

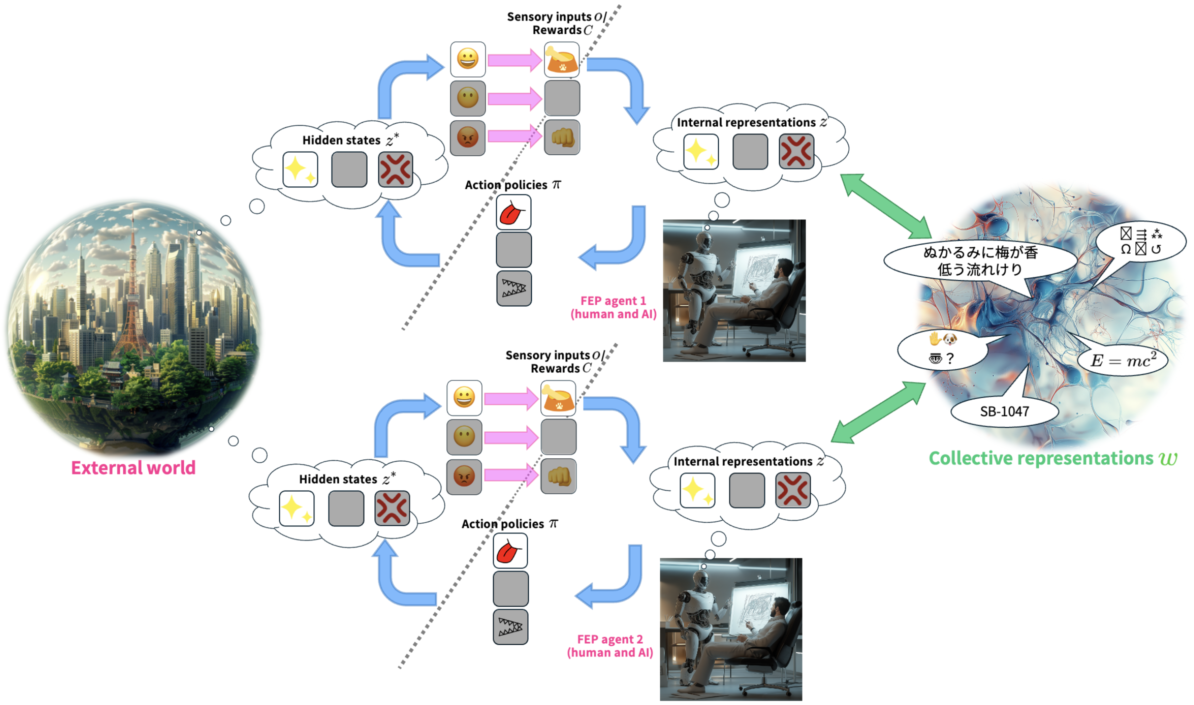
Sensory inputs o / Rewards C
Hidden states z^*
Internal representations z
Action policies π
FEP agent 1 (human and AI)
FEP agent 2 (human and AI)
External world
ぬかるみに梅が香
低う流れけり
Ω ↺
✋🐶
?
$E = mc^2$
SB-1047
Collective representations w

Figure 1. **A schematic illustrating Collective Predictive Coding.** In CPC, each agent updates its own internal states (latent variables) while also interacting with shared external representations (language, symbols, etc.). The difference from standard Predictive Coding is the *collective regularization term*, allowing alignment of agents through the environment's symbolic layer. This collective layer effectively acts like an additional control force in the system's stochastic dynamics.

3.1. Bayesian Updating and Stochastic Gradient Descent

Consider a target distribution (posterior) of parameters x:

$$\pi(x) \ \propto \ \exp\big[-U(x)\big], \tag{3}$$

where $U(x)$ represents the negative log of the (unnormalized) posterior plus any regularizers. In PC or CPC, $U(x)$ typically corresponds to the expected negative log-likelihood plus KL divergences appearing in the free energy. Minimizing $U(x)$ (or equivalently, the free energy) leads to posterior-consistent estimates.

When we allow stochasticity to preserve exploration—instead of pure gradient descent—we obtain a **Langevin** update:

$$\frac{\mathrm{d}x}{\mathrm{d}t} \ = \ -\nabla_x U(x) \ + \ \xi(t), \tag{4}$$

where $\xi(t)$ is typically Gaussian noise. Under mild assumptions (e.g., Gaussian white noise, sufficiently small step size, and smooth $U(x)$), the long-term behavior of $x(t)$ samples from $\pi(x)$.

3.2. Fokker–Planck Equation and Stationary Distribution

The distribution $\mu(x,t)$ of states x at time t in the SDE Equation (4) evolves according to the *Fokker–Planck equation*:

$$\frac{\partial}{\partial t}\,\mu(x,t) \ = \ \nabla_x \cdot \Big(\nabla_x U(x)\,\mu(x,t)\Big) + \Delta_x\,\mu(x,t), \tag{5}$$

where Δ_x is the Laplacian operator in x-space (assuming isotropic diffusion). The key fact is:

- **Stationary Distribution:** If $\mu(x,t)$ converges to a stationary solution $\pi(x)$ such that $\frac{\partial}{\partial t}\pi(x) = 0$, then one can show

  $$\pi(x) \ \propto \ \exp[-U(x)],$$

  matching the target posterior. Essentially, the drift term $-\nabla_x U(x)$ in Equation (4) drives the system toward the posterior distribution, while the diffusion (noise) ensures exploration.

Implication for Bayesian Inference. Thus, *the same distribution* that one obtains via Bayesian updates (i.e., $\pi(x) \propto e^{-U(x)}$) emerges as the stationary solution of the Langevin dynamics in Equation (4). This is the formal reason why substituting Bayesian updates with a Langevin process is valid under typical conditions (small steps, Gaussian noise, etc.).

For **Predictive Coding**, $U(x)$ incorporates the expected surprise and regularization terms in F_{PC}, while for **Collective**

Predictive Coding, $U(x)$ includes additional coupling from w. In both cases, the Fokker–Planck perspective demonstrates that the dynamics converge to the corresponding posterior distribution.

3.3. Conclusion: Linking Free Energy to Stochastic Dynamics

Minimizing the free energy in a deterministic gradient fashion might yield a point estimate. However, introducing noise (leading to the Langevin SDE) can maintain a distributional perspective, sampling around modes of the free-energy landscape in proportion to $e^{-U(x)}$. Consequently, analyzing PC or CPC in terms of Langevin dynamics is justified by the Fokker–Planck equation's guarantee that the equilibrium distribution remains the intended Bayesian posterior.

4. Emergence of an Additional Force in CPC

We now apply the Fokker–Planck and Langevin reasoning to compare the updates of PC and CPC. The key difference is the extra *collective* term $R(x)$ from the shared representation w, yielding a new force in the system's dynamical equations.

4.1. PC: Baseline Langevin Dynamics

In standard PC,

$$U_{\mathrm{PC}}(x) \;\approx\; -\ln p(\mathbf{o} \mid x) \;+\; \text{KL regularizers},$$

leading to the Langevin equation:

$$\frac{\mathrm{d}x}{\mathrm{d}t} = -\nabla_x U_{\mathrm{PC}}(x) \;+\; \xi(t). \tag{6}$$

By Equation (5), the stationary distribution of x is $\pi_{\mathrm{PC}}(x) \propto e^{-U_{\mathrm{PC}}(x)}$, mirroring the Bayesian posterior for the generative model.

4.2. CPC: Augmented Potential and External Force

In Collective Predictive Coding, the effective energy contains an extra piece:

$$U_{\mathrm{CPC}}(x) \;=\; U_{\mathrm{PC}}(x) \;+\; R(x),$$

where $R(x)$ encapsulates the dependence on the shared representation w and the collective regularization. The associated Langevin equation is:

$$\frac{\mathrm{d}x}{\mathrm{d}t} = -\,\nabla_x \Big[U_{\mathrm{PC}}(x) \;+\; R(x) \Big] \;+\; \xi(t). \tag{7}$$

Thus, the gradient $\nabla_x R(x)$ modifies each agent's dynamics. Physically, one can interpret $-\nabla_x R(x)$ as an *external force* that arises only because of the multi-agent coupling through w.

Why Does This Matter? Even if an individual agent's local observations are fully explained (i.e., it has minimal individual free energy), the additional term $R(x)$ can still push that agent to change its state in order to better synchronize with the shared representation. This mechanism underlies CPC's capacity for *collective alignment*, in which language, symbols, or shared knowledge structures impose a "soft" constraint on each agent's updates.

Fokker–Planck Equilibration for CPC. Applying the same Fokker–Planck argument, the equilibrium distribution for Equation (7) is $\pi_{\mathrm{CPC}}(x) \propto e^{-U_{\mathrm{PC}}(x)-R(x)}$. In other words, the distribution at stationarity is shaped by both the local predictive-coding objective *and* the collective coupling induced by w.

5. Discussion

We have shown that CPC introduces an external force via an extra potential term in the Langevin equations, and that the Fokker–Planck equilibrium matches the posterior distribution implied by CPC's generative model. Here, we expand on broader implications, limitations, and future directions.

5.1. Hybrid Human–AI Coordination

In scenarios where humans and AI agents share a symbolic framework (e.g., natural language instructions, code repositories, or knowledge graphs), the external force $-\nabla_x R(x)$ can be interpreted as:

> *A steering mechanism by which updates to the shared representation w shape the entire system's evolution.*

Rather than micromanaging each agent's internal states, system designers or stakeholders can influence w—for instance, by embedding certain norms or goals. The CPC free energy then ensures that all agents shift their beliefs/strategies accordingly, as long as they remain within the CPC paradigm.

5.2. Practical Caveats and Imperfections

Real-world conditions may undermine the neat Fokker–Planck-based story:

- **Communication Noise:** Agents might only partially receive updates to w, or interpret them differently (especially human vs. AI).
- **Symbol Emergence:** The space of possible w could expand over time, with new symbols spontaneously created. Modeling $p_\theta(w)$ thus becomes nontrivial.
- **Incentive Misalignment:** Agents may not purely minimize the CPC free energy if they have external rewards

or strategic motives that deviate from it.

- **Computation Limits:** In high-dimensional or continuous domains, approximating $\nabla_x R(x)$ may be computationally expensive.

Hence, while the theory is conceptually compelling, bridging to real applications calls for robust approximations and possibly hierarchical or modular expansions of CPC.

5.3. Ethical and Safety Considerations

From an AI safety perspective, CPC's external force can be a double-edged sword:

- **Positive Aspect:** One can embed norms, ethics, or alignment constraints into w, so that CPC agents remain consistent with human values.
- **Negative Aspect:** If w is hijacked or corrupted (e.g., maliciously introduced misinformation), the entire system could be pulled in undesirable directions.

Designers must therefore ensure the shared representation is secure, interpretable, and open to scrutiny.

5.4. Open Challenges and Future Work

(1) Empirical Validation in Multi-Agent Systems. We suggest testing CPC in multi-agent reinforcement learning environments where agents pass discrete or continuous symbols, then verifying whether the learned "external force" indeed fosters alignment or coordination.

(2) Human–AI Co-learning. Another direction is to incorporate humans-in-the-loop to see how well they adapt to AI-suggested symbols and vice versa. This might illuminate emergent communication phenomena and symbol grounding issues.

(3) Generalizing w. In some applications, w might itself be structured (e.g., a graph of scientific concepts) or unstructured (e.g., text corpora). Different representations may yield different forms of $R(x)$ and thereby distinct modes of collective control.

6. Conclusion

In this paper, we have provided a comprehensive account of how *Collective Predictive Coding* (CPC) differs from conventional *Predictive Coding* (PC). We reintroduced the detailed free-energy components in both frameworks, emphasized how Bayesian updating maps to Langevin equations, and then used the Fokker–Planck perspective to clarify that the stationary distribution of these SDEs coincides with the respective Bayesian posteriors for PC and CPC.

Key Takeaways:

- PC explains how individual agents minimize free energy by aligning their latent states with observations and priors.
- CPC incorporates a shared external representation w, adding a collective term to the free energy that manifests as an external force in the Langevin dynamics.
- The Fokker–Planck framework ensures that substituting Bayesian updates with Langevin equations does not distort the final equilibrium distribution, providing formal grounding for the approach.
- Real-world applications involve challenges like communication noise, emergent symbols, and partial alignment, but CPC suggests a promising path to orchestrate hybrid human–AI systems through a common symbolic layer.

Future Directions. Beyond the immediate theoretical extensions, we encourage work on:

1. **Robust Simulation Studies:** To empirically test how well CPC's external force fosters or hinders coordination in noisy, multi-agent contexts.
2. **Symbol Emergence Modeling:** Investigating how w evolves in open-ended environments and how new symbols might dynamically reshape $R(x)$.
3. **Safe and Accountable Implementations:** Ensuring that any externally controlled representation remains transparent and not vulnerable to malicious interference.

By showing in detail why and how CPC aligns with Bayesian–Langevin theory, this manuscript hopefully clarifies the mechanisms driving the additional collective potential and inspires further research into safe, coordinated multi-agent systems.

Impact Statement

This work refines the conceptual and mathematical link between Bayesian inference, Predictive Coding, and Collective Predictive Coding. By highlighting how shared representations can introduce a soft external force in multi-agent systems, we underscore both the benefits (easier alignment and collective control) and risks (susceptibility to manipulation or corruption). We believe clearer theoretical grounding of CPC can stimulate responsible innovation in cooperative AI and human–AI teaming, while also cautioning that the power of shared symbolic channels should be carefully governed.

References

Friston, K. and Kiebel, S. Predictive coding under the free-energy principle. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1521):1211–1221, May 2009. doi: 10.1098/rstb.2008.0300.

Rao, R. P. N. and Ballard, D. H. Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2: 79–87, 1999.

Taniguchi, T. Collective predictive coding hypothesis: Symbol emergence as decentralized Bayesian inference. *Frontiers in Robotics and AI*, 11:1353870, 2024.