Concept Bottleneck Model with Emergent Communication Framework for Explainable AI*

Farnoosh Javar[1], Kei Wakabayashi[1,*]

[1]University of Tsukuba, Ibaraki, Japan

Abstract

Interpretable machine learning seeks to enhance transparency in model decision-making, particularly in high-stakes applications. Concept bottleneck models (CBMs) improve interpretability by using human-defined concepts as intermediate representations. Yet, they often depend on extensive manual annotations and may fail to capture relevant features beyond predefined concepts. We propose an iterative communication emergence framework for interpretable machine learning that integrates a concept bottleneck model with data-driven discovery of latent features. Our approach employs a sender–receiver architecture, where the sender encodes raw inputs into discrete latent signals refined via reinforcement learning, and the receiver uses these latent concepts to predict outcomes. Latent representations are aligned post hoc with human-observable concepts, which are automatically generated by a language model and validated statistically, enabling transparent explanations while reducing reliance on manual annotations. Experiments on a cat breed classification task demonstrate that our framework maintains high predictive performance while progressively refining interpretable concept representations. Results suggest that emergent latent concepts can meaningfully align with human-understandable attributes, facilitating more flexible and scalable interpretability in deep learning models.

Keywords

Concept bottleneck models, Emergent communication, Multi-agent communication

1. Introduction

Interpretable machine learning is essential for high-stakes decision-making, where understanding model behavior is critical. Rather than relying on post-hoc explanations of black-box models, researchers advocate for inherently interpretable models [1]. Large language models (LLMs) increase interpretability challenges, particularly in specialized domains such as health care and justice, where transparency and domain expertise are crucial. Their opaque reasoning process undermines trust in AI-driven decision-making, necessitating frameworks that enhance both accuracy and interpretability [2, 3].

Concept bottleneck models provide a structured approach to interpretability by introducing an intermediate layer of human-defined concepts, enabling direct user intervention. These models achieve competitive performance while enhancing transparency by allowing users to modify concept values. However, CBMs are constrained by their reliance on manual concept annotations, which can be costly or infeasible in specific domains. Additionally, they often underperform compared to unconstrained models, limiting their practical adoption [4]. CBMs also limit model flexibility by enforcing predictions based solely on predefined concepts, restricting the model's ability to capture latent concepts that may exist in the data but are not explicitly defined. Additionally, soft concept predictions can introduce information leakage, where intermediate representations inadvertently encode task-specific information beyond the intended concepts, reducing interpretability. Balancing predictive performance and transparency continues to pose a central challenge in the field of interpretable AI [5].

To address these limitations, post-hoc concept bottleneck models (PCBMs) have been proposed to transform pre-trained neural networks into CBMs without sacrificing performance. PCBMs facilitate external concept integration and enable efficient model editing to mitigate dataset biases [6]. Another

Late-breaking work, Demos and Doctoral Consortium, colocated with the 3rd World Conference on eXplainable Artificial Intelligence: July 09–11, 2025, Istanbul, Turkey

*Corresponding author.

s2326091@u.tsukuba.ac.jp (F. Javar); kwakaba@slis.tsukuba.ac.jp (K. Wakabayashi)

0009-0008-1050-2019 (F. Javar); 0000-0001-6898-4833 (K. Wakabayashi)

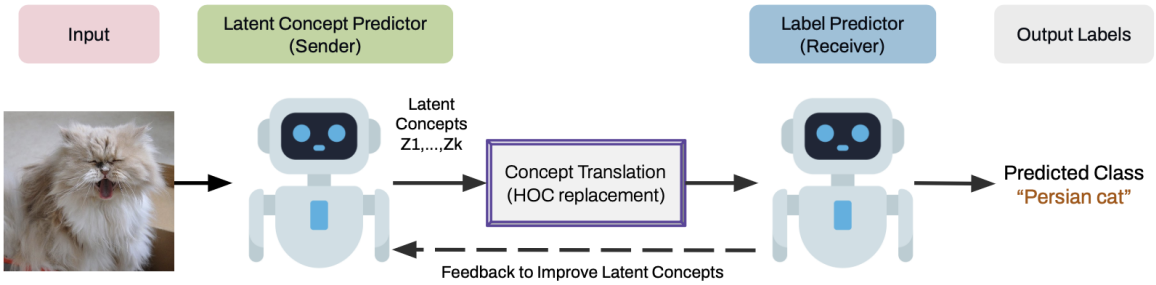


Figure 1: Overview of the proposed framework. The sender learns latent concepts, the receiver makes predictions, and a reinforcement learning loop iteratively refines the concept representations.

promising approach involves automatically discovering latent concepts from model representations, reducing reliance on predefined annotations while maintaining interpretability and narrowing the accuracy gap between interpretable and black-box models [7].

Recent advancements leverage LLMs to generate human-readable concepts automatically. For example, the language in a bottle (LaBo) framework utilizes GPT-3 to replace manual concept annotations with natural language descriptions. However, this approach introduces challenges such as assessing the validity of generated concepts and mitigating information leakage [8]. Additionally, recent research in emergent multi-agent communication investigates how AI agents develop internal languages to coordinate and solve tasks collaboratively. These emergent languages can encode task-relevant abstractions that are not explicitly programmed, potentially revealing novel representations. However, aligning these representations with human semantics remains an open challenge, as the symbols and structures developed by AI agents may not directly correspond to human-understandable concepts, requiring further interpretability methods [9]. In this work, we propose a discrete emergent language framework that integrates aspects of concept bottlenecks with data-driven discovery of latent features. Instead of relying solely on a fixed set of human-defined concepts, our approach enables a sender–receiver pair of modules to develop a symbolic communication system through iterative training, subject to a discrete bottleneck. The language is considered emergent because the meanings of these symbols are not explicitly predefined by humans; instead, they develop as the system optimizes sender–receiver interactions for the given task. By constraining communication to a discrete channel, the model's internal reasoning is structured around human-interpretable latent concepts when alignment with known human concepts occurs. When a learned latent concept does not correspond to any predefined human concept, it is retained as a potentially novel feature. These latent factors remain part of the communication protocol and can later be analyzed by domain experts to assess their relevance or meaning. Our framework seeks to balance predictive performance with interpretability, allowing models to capture meaningful latent patterns while mitigating the limitations of predefined concept bottlenecks.

2. Proposed Approach

We propose an iterative language emergence framework within a concept bottleneck model pipeline to achieve interpretable and scalable AI. The framework consists of two cooperative agents—a sender and a receiver—that communicate through discrete signals of latent concepts. The sender converts raw inputs into binary latent concept vectors, while the receiver, implemented as a simple linear classifier, maps these vectors directly to predicted labels. A reinforcement learning-based feedback loop using proximal policy optimization (PPO) [10] iteratively refines these representations to improve both accuracy and interpretability (see Fig. 1).

2.1. Problem Definition

We address an image classification task with training dataset $\{(x^{(i)}, y^{(i)})\}_{i=1}^{N}$ where $x^{(i)}$ is an image, $y^{(i)}$ is the associated class label, and N denotes the training dataset size. In addition, we have access to a set of

human-observable concepts (HOCs) that can provide interpretable annotations (e.g., "slender body" or "no visible whiskers"). Let $D = \{d_1, \dots, d_M\}$ be the set of natural language descriptions of the HOCs, and $HOC^{(i)} = \{HOC_1^{(i)}, \dots, HOC_M^{(i)}\}$ be binary values indicating the presence or absence of the m-th HOC in the i-th image. We assume that LLMs can be leveraged to automatically generate candidate HOCs D and $HOC_j^{(i)}$ from class descriptions, domain-specific glossaries, or other textual corpora by prompting with questions such as "What are the visual features of y?'' (where y is the class label) and "does the image $x^{(i)}$ have the visual feature d_j?", respectively. In contrast to traditional CBMs, this assumption does not require concept generation by human domain experts. We emphasize that, however, the available HOCs are not necessarily effective features for the classification task.

Our goal is to develop a model that classifies images accurately *and* offers transparent reasoning based on a combination of *discrete* interpretable concepts. The discrete binary representation of HOCs (i.e., hard concepts) enhances interpretability and prevents information leakage [5]. Effectiveness is evaluated along two axes: classification accuracy and the extent to which human supervision—particularly in the form of HOC annotations—is required at test time. Reducing this annotation burden, even when HOCs are initially bootstrapped via LLMs, is desirable for scaling interpretable AI systems to new instances.

2.2. Sender–Receiver Framework and Emergent Communication Protocol

The sender is a feedforward neural network (FNN) π_θ parameterized by θ, which processes raw input features $x^{(i)}$ (e.g., an image of a cat) and outputs a vector of K discrete latent concepts, $Z^{(i)} = (Z_1^{(i)}, \dots, Z_K^{(i)})$. These latent concepts represent task-relevant patterns that emerge from data. To enforce discreteness, the continuous outputs of the FNN are thresholded using a predefined threshold τ:

$$Z_k^{(i)} = \begin{cases} 1, & \text{if } \pi_\theta(Z_k^{(i)} \mid x^{(i)}) > \tau, \\ 0, & \text{otherwise,} \end{cases} \tag{1}$$

where $\pi_\theta(Z_k^{(i)} \mid x^{(i)})$ is the predicted probability that concept $Z_k^{(i)}$ is active for a given input $x^{(i)}$. This binarization yields a vector (e.g., $Z^{(i)} = [1, 0, 0, 1, 1]$), which serves as the message sent to the receiver.

The set of discrete latent concept symbols forms a communication protocol between the agents. Each Z_k can be seen as a "word" in the emergent communication that the sender uses to describe the input's relevant properties.

The receiver agent P_ϕ, parameterized by ϕ, is a simple linear classifier that takes the sender's message $Z^{(i)}$ as input and outputs a prediction $\hat{y}^{(i)}$ of the final class label (e.g., the breed of a cat, such as "Persian"). The receiver is essentially the label predictor in the CBM pipeline: it maps from latent concepts to the target output. Because we ensure that any information used by the receiver to make the prediction must pass through the bottleneck of discrete concepts (intended to be human-understandable), its decision process is more transparent than a direct end-to-end model.

We optimize the sender's policy to communicate useful concepts using reinforcement learning (RL). The sender is updated with a policy gradient method – specifically, proximal policy optimization (PPO) – to maximize the expected reward. The reward for the sender's action (i.e., message $Z^{(i)}$) is computed after the receiver predicts a label $\hat{y}^{(i)}$ as follows:

$$R = \begin{cases} +\alpha \cdot \max(\hat{y}), & \text{if } \arg\max(\hat{y}) = y, \\ -\beta \cdot \max(\hat{y}), & \text{otherwise,} \end{cases} \tag{2}$$

where $\max(\hat{y})$ represents the maximum logit for the predicted class, α is a positive scaling factor for correct predictions, and β controls the penalty for incorrect classifications. At each training iteration, we sample a batch of inputs, let the sender produce messages Z, and let the receiver predict labels $\hat{y}$. Then, the receiver's parameter is trained via supervised learning using cross-entropy loss, and the sender's parameter θ is updated by the PPO algorithm. This training process can be seen as the agents jointly evolving a more effective shared communication protocol.

2.3. Aligning Latent Concepts with Human-Observable Concepts

While the latent concepts emerge autonomously during training, we align them with HOCs in a post hoc fashion using statistical validation. The proposed algorithm consists of N_{cycle} iterative cycles, where the t-th cycle consists of three steps: (1) training the receiver for N_{epochs} using supervised learning and updating the sender via PPO during each epoch, (2) choosing the best parameter pair $(\theta^{(t)}, \phi^{(t)})$ among the N_{epochs} epochs based on validation accuracy, (3) establishing a new alignment mapping from a latent concept to a HOC.

In the step (3), the proposed method generates the latent concept vector $Z^{(i)}$ from the sender $\pi_{\theta^{(t)}}$ for each image $x^{(i)}$ in the training set. For each pair of latent concept Z_k and HOC dimension HOC_j, we construct a contingency table of co-occurrence statistics between the presence vectors $\{Z_k^{(1)}, \ldots, Z_k^{(N)}\}$ and $\{HOC_j^{(1)}, \ldots, HOC_j^{(N)}\}$ and apply Fisher's exact test to calculate the p-value of the dependence. We take the pair (k, j) such that the p-value is lowest and establish a mapping from the latent concept Z_k to HOC_j if it indicates a statistical significance (i.e., $p < 0.1$).

Once a mapping from Z_k to HOC_j is established, we substitute the value of Z_k with the corresponding HOC annotation HOC_j in all subsequent processes without modifying the model's internal computation or predictions. For example, consider a scenario where the agents are trained to classify images of cats, and Z_2 is aligned with the HOC_1 "long whiskers" at the end of the first cycle. If the sender outputs $Z^{(i)} = [1, 1, 0, 1, 1]$ for an image $x^{(i)}$ of cat without long whiskers, $Z_2^{(i)}$ is replaced with $\mathrm{HOC}_1^{(i)} = 0$, and the receiver will receive the modified vector $[1, 0, 0, 1, 1]$ during subsequent training and evaluation. This mapping allows the model's decisions to be interpreted in terms of known concepts when available, or left as abstract latent factors when no significant alignment is found.

The alignment mapping from latent concepts to HOCs is cumulative; once a latent concept is aligned to a HOC, that mapping is retained in all subsequent cycles. Suppose that Z_5 is aligned with HOC_8 "flat nose" in the next cycle of the example above. If the sender outputs $Z^{(i)} = [1, 1, 0, 0, 1]$ for an image of a cat with a flat nose and no long whiskers ($\mathrm{HOC}_1 = 0$ and $\mathrm{HOC}_8 = 1$), the modified vector $[1, 0, 0, 0, 1]$ will be sent to the receiver. The receiver predicts the class label, now with a more interpretable explanation. This refinement process accumulates aligned concepts while preserving predictive performance.

Over successive iterations, this cycle progressively refines both the predictive power and interpretability of the model. Latent concepts that prove useful for the task and show consistent alignment with HOCs become more semantically meaningful, while unaligned concepts may remain the model's newly discovered factors as *novel concepts*, which can be studied further by experts or potentially added to the set of HOCs for future iterations. Importantly, the sender's parameters remain persistent across iterations, enabling stable concept refinement, while the alignment mapping evolves to capture emerging associations. This design allows the model to maintain high task accuracy while offering increasingly transparent, concept-based explanations for its decisions.

3. Experiments

3.1. Dataset

We evaluate our framework on a subset of the Oxford-IIIT-Pet dataset [11] containing four cat breeds: Ragdoll, Persian, Sphynx, and Russian Blue. ResNet50-extracted features are used as input representations. To avoid noise from LLM-generated concepts, we adopt a two-step verification: candidate HOCs are first generated via generative AI and then manually annotated for reliability. This setup enables robust alignment between latent concepts and human-observable attributes, facilitating rigorous evaluation of our method. The final dataset comprises 400 images, each annotated with 26 binary HOCs ($M = 26$), and is divided into training ($N = 320$), validation, and test sets using an 80/10/10 stratified split.

Table 1
Latent concept alignments and model validation performance across training iterations. Alignments are validated using Fisher's exact test ($p < 0.1$). Each aligned HOC is shown with its natural language description.

Iteration	Latent Concept	Aligned HOC	p-value	Val. Acc.	Test Acc.
1	Z_{15}	HOC_{21} (Slender body)	4.49×10^{-3}	85%	75%
2	Z_5	HOC_{24} (Slim oval paws)	1.46×10^{-3}	95%	85%
3	Z_{17}	HOC_{16} (No visible whiskers)	6.53×10^{-3}	92.5%	75%
4	Z_{24}	HOC_{23} (Sturdy round paws)	1.84×10^{-3}	90%	80%
5	Z_{16}	HOC_{18} (Long, tapered tail)	8.03×10^{-3}	97.5%	100%

3.2. Experimental Configuration

Our experiments follow the iterative training procedure described in Section 2. The sender and receiver are trained over five iterative cycles ($N_{cycle} = 5$), each consisting of 30 epochs ($N_{epochs} = 30$), with a batch size of 32. PPO settings include n_steps = 64, a batch size of 16, and a learning rate of 3×10^{-4}. All other hyperparameters follow the default values in stable-baselines3 (v2.5.0).

The alignment between latent concepts and HOCs is updated at the end of each iterative cycle using Fisher's exact test, applying a significance threshold of $p < 0.1$. Model performance is evaluated via classification accuracy and cross-entropy loss on both validation and test sets. We also provide qualitative analysis by inspecting the translated, human-readable concept vectors.

To evaluate the framework's full potential under controlled conditions, we isolate the model's behavior from noise introduced by LLMs. While candidate HOCs are automatically generated via a large language model, human involvement is restricted to annotating the presence or absence (0/1) of each predefined concept. This preserves the automation of concept discovery while ensuring label quality.

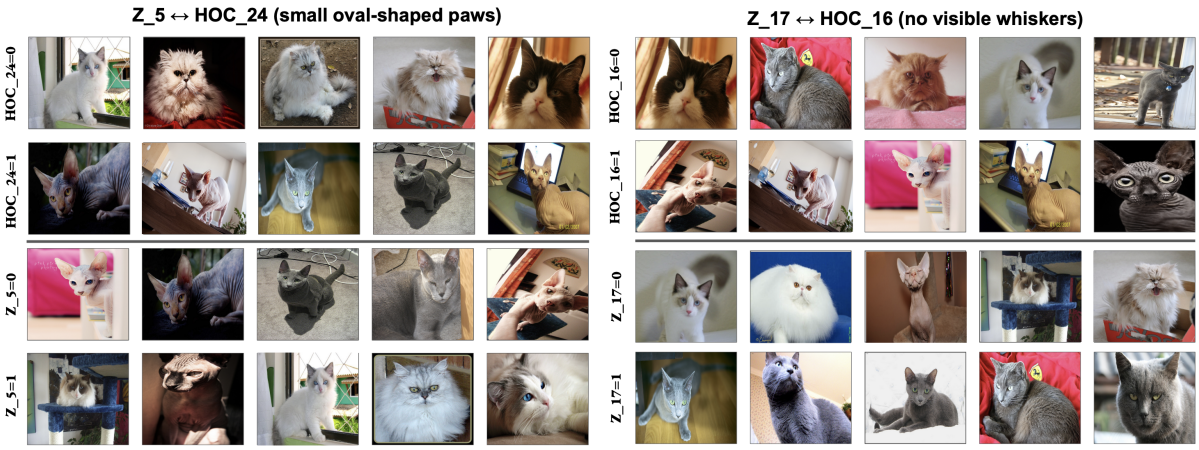
For comparison, we trained a traditional CBM using the same label predictor architecture as our proposed method. To evaluate the effect of concept supervision, we trained CBMs using 5, 10, and 15 randomly selected HOCs. Each model was trained for 150 ($= N_{cycle} \times N_{epochs}$) epochs using a combined loss function: binary cross-entropy for HOC prediction and cross-entropy for label prediction. We repeated each configuration across five random seeds and report the mean test accuracy.

3.3. Experimental Results and Discussion

We report results from a single trial of the proposed iterative framework, which progressively aligned latent concepts with HOCs over five training iterations. In each iteration, one statistically significant alignment was identified via Fisher's exact test ($p < 0.1$)—e.g., Z_{15} with HOC_{21} (slender body) in Iteration 1 and Z_{16} with HOC_{18} (long, tapered tail) in Iteration 5. Full alignments and metrics are shown in Table 1. These results indicate that the framework refines latent structures over time, aligning emergent concepts with HOCs while maintaining high predictive accuracy. Unlike traditional CBMs that rely on predefined concepts, our method identifies meaningful associations post hoc, enabling more flexible and scalable interpretability.

Validation accuracy increased from 85% to 97.5%, and test accuracy from 75% to 100% over iterations. Validation loss declined from 0.8490 in iteration 1 to 0.3675 in iteration 5, with minor fluctuations across epochs. These metrics indicate improved performance over time, although gains were not strictly monotonic. This suggests a potential relationship between concept alignment and model confidence. As training progressed, more stable predictions were observed—likely enabled by PPO-based sender-receiver optimization, which allowed latent representations to evolve without causing abrupt changes in downstream accuracy. However, whether such stability generalizes to datasets with more complex feature distributions remains an open question.

To further examine the relationship between latent units and human-interpretable concepts, we visualize sample groupings in Figure 2. Subfigure 2a compares latent concept Z_5 with HOC_{24}, and



(a) Z_5 and HOC_{24} (Slim oval paws). (b) Z_{17} and HOC_{16} (no visible whiskers).

Figure 2: Visual correspondence between latent concepts and human-interpretable attributes. (a) Latent concept Z_5 compared with HOC_{24} (Slim oval paws). (b) Latent concept Z_{17} compared with HOC_{16} (no visible whiskers). Each panel shows samples grouped by ground-truth concept (top) and latent activation (bottom). The samples shown are randomly selected from each group.

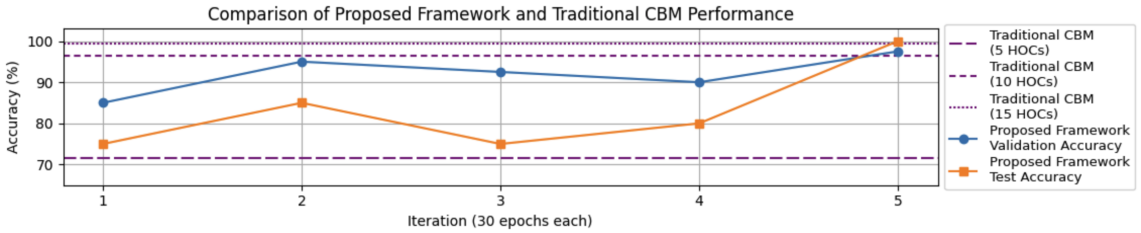


Figure 3: Validation and test accuracy across iterations of our proposed framework, compared to the mean test accuracy of the traditional CBM. Each iteration comprises 30 training epochs, totaling 150 epochs. The traditional CBM was trained for the same duration without iterative refinement.

Subfigure 2b compares Z_{17} with HOC_{16}. In each panel, the top two rows are grouped by HOC values, and the bottom two by latent activations. For Z_5, some visual consistency is observable with the paw shape associated with HOC_{24}. For Z_{17}, activated samples tend to belong to a consistent breed (e.g., Russian Blue), though the whisker attribute associated with HOC_{16} is not always present. We discuss this type of partial alignment and its implications in Section 4. These examples reflect the post hoc nature of alignment in our framework—some latent features align well with visual attributes, while others only partially capture the associated HOC. This underscores the need for more robust validation methods to assess the semantic coherence of emergent representations.

The traditional CBM achieved a mean test accuracy of 71.5% when trained with 5 randomly selected HOCs. Increasing the number of HOCs to 10 and 15 improved the mean test accuracy to 96.5% and 99.5%, respectively. These results indicate that strong performance in the traditional CBM setting is dependent on supervision from a sufficiently large and informative concept set. In contrast, our framework achieved 100% test accuracy by the final iteration while relying on only 5 HOC alignments selected through iterative discovery. Performance trends are illustrated in Figure 3. This highlights its ability to dynamically refine concept representations based on data, without requiring full concept annotations. The progressive nature of alignment, combined with strong classification performance, suggests potential advantages in generalization and interpretability when compared to fixed, manually labeled bottlenecks.

4. Limitations and Future Directions

While the iterative communication-emergence CBM framework shows promise, it has limitations in concept coverage and consistency. In our experiments, only five latent concepts aligned confidently with HOCs, leaving many unmapped due to weak associations or missing human-defined attributes. Future work could cluster activations and analyze them via language models or human feedback to expand or refine the HOC space. Some unmapped concepts may reflect novel patterns, but not all are necessarily interpretable.

We did not explicitly assess mapping consistency across training runs or datasets. Although aligned concepts remained fixed once identified during our single trial, their stability under different data distributions or longer training remains unverified. Initial assignments may vary due to dataset properties or learning dynamics, potentially leading to concept drift. Addressing this may require more flexible alignment strategies and evaluations across multiple runs [12, 13].

Limitations also arise in the scope of interpretability. Our evaluation relies on statistical correlations with predefined HOCs, confirming whether a latent factor rediscovers a known concept, but not how it is internally represented. The fixed binarization threshold ($\tau = 0.5$) may also affect which concepts align but was not evaluated. Additionally, the sender is optimized for prediction accuracy, not concept-level supervision, allowing spurious correlations to persist. Although aligned variables are filtered before prediction, the sender may still generate them, with no mechanism to discourage reliance on non-interpretable factors. Future work could examine the threshold's effect, use visualization or attribution methods to interpret latent factors, and incorporate causal or adversarial regularization to mitigate spurious patterns. Expert review may help define new HOCs, and applying the framework to diverse datasets will be essential to assess generalizability. The traditional CBM was not further fine-tuned. While the comparison still highlights differences in reliance on supervision, future work should explore whether tuning CBM hyperparameters improves performance. In our PPO reward, we use the maximum predicted logit as a confidence measure. While this supports interpretability, exploring alternatives such as maximum softmax probability or output entropy remains a promising direction.

A key limitation is that the current alignment strategy permits only many-to-one mappings from latent concepts to HOCs, without supporting many-to-many relationships. In practice, multiple latent concepts may redundantly align with the same HOC, while a single latent concept may also encode multiple abstract or overlapping attributes. Post hoc alignment does not influence training in our current setup. Incorporating soft alignment rewards could encourage interpretability as associations emerge, but must be balanced against model flexibility to avoid the rigidity of traditional CBMs.

5. Conclusion

This work presents an iterative communication emergence framework for interpretable machine learning, integrating reinforcement learning into a CBM to enable emergent concept discovery. Unlike traditional CBMs that rely on predefined annotations, our approach allows the model to autonomously align latent representations with human-understandable attributes through iterative refinement.

Experiments on cat breed classification demonstrate that interpretability improves progressively over training—latent concepts become increasingly aligned with observable features—while predictive accuracy also improves. The model achieved perfect test performance by the final iteration, despite relying on only five post hoc-aligned HOCs, without direct concept supervision. Aligned concepts remained stable once discovered, suggesting robustness in alignment.

While our framework enables latent concepts to align with human-interpretable attributes, it has limitations. Many concepts remain unmapped, and the current many-to-one alignment may oversimplify complex or overlapping patterns. Since alignment occurs post hoc, it does not guide training. Future work could incorporate soft alignment objectives, adaptive mappings, and visualization techniques to better integrate interpretability into the learning process. Ultimately, bridging emergent representations with human semantics offers a path toward more transparent and trustworthy AI systems.

Declaration on Generative AI

The author(s) used GPT-4o and Grammarly for grammar and minor wording improvements and reviewed and edited the content as needed. The author(s) take full responsibility for the publication's content.

Acknowledgements

This work was partly supported by JST CREST grant number JPMJCR22M2.

References

[1] C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, Nature Machine Intelligence 1 (2019) 206 – 215.

[2] H. Luo, L. Specia, From understanding to utilization: A survey on explainability for large language models, arXiv preprint (2024). URL: https://arxiv.org/abs/2401.12874. arXiv:2401.12874.

[3] C. Singh, J. P. Inala, M. Galley, R. Caruana, J. Gao, Rethinking interpretability in the era of large language models, arXiv preprint (2024). URL: https://arxiv.org/abs/2402.01761. arXiv:2402.01761.

[4] P. W. Koh, T. Nguyen, Y. S. Tang, S. Mussmann, E. Pierson, B. Kim, P. Liang, Concept bottleneck models, in: H. D. III, A. Singh (Eds.), Proceedings of the 37th International Conference on Machine Learning, volume 119 of *Proceedings of Machine Learning Research*, 2020, pp. 5338–5348.

[5] M. Havasi, S. Parbhoo, F. Doshi-Velez, Addressing leakage in concept bottleneck models, in: Advances in Neural Information Processing Systems, volume 35, 2022, pp. 23386–23397.

[6] M. Yuksekgonul, M. Wang, J. Y. Zou, Post-hoc concept bottleneck models, in: Proceedings of the International Conference on Learning Representations, 2023.

[7] S. Schrodi, J. Schur, M. Argus, T. Brox, Concept bottleneck models without predefined concepts, arXiv preprint (2024). URL: https://arxiv.org/abs/2407.03921. doi:10.48550/arXiv.2407.03921. arXiv:2407.03921.

[8] Y. Yang, A. Panagopoulou, S. Zhou, D. Jin, C. Callison-Burch, M. Yatskar, Language in a bottle: Language model guided concept bottlenecks for interpretable image classification, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 19187–19197. doi:10.1109/CVPR52729.2023.01839.

[9] A. Lazaridou, M. Baroni, Emergent multi-agent communication in the deep learning era, arXiv preprint (2020). URL: https://arxiv.org/abs/2006.02419. arXiv:2006.02419.

[10] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, O. Klimov, Proximal policy optimization algorithms, arXiv preprint (2017). URL: https://arxiv.org/abs/1707.06347. arXiv:1707.06347.

[11] O. M. Parkhi, A. Vedaldi, A. Zisserman, C. V. Jawahar, Cats and dogs, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 3498–3505. doi:10.1109/CVPR.2012.6248092.

[12] R. Chaabouni, E. Kharitonov, D. Bouchacourt, E. Dupoux, M. Baroni, Compositionality and generalization in emergent languages, in: Proceedings of the Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2020, pp. 4427–4442. URL: https://aclanthology.org/2020.acl-main.407. doi:10.18653/v1/2020.acl-main.407.

[13] M. Rita, C. Tallec, P. Michel, J.-B. Grill, O. Pietquin, E. Dupoux, F. Strub, Emergent communication: Generalization and overfitting in lewis games, in: Advances in Neural Information Processing Systems, volume 35 of *NIPS '22*, Curran Associates Inc., 2022, pp. 16744–16760.

A. Online Resources

The source code for our experiments is available at:

- https://github.com/tzkwkblab/cbm-emergent-communication