CO-ADAPTIVE EXPLANATION INTERFACES: ALIGNING AI AND HUMAN REASONING THROUGH DUAL-CHANNEL FEEDBACK

Anonymous authors
Paper under double-blind review

ABSTRACT

We introduce Co-Adaptive Explanation Interfaces, an interactive XAI framework that models individual users' cognitive biases and delivers dual-channel explanations: (1) content justifications for model decisions and (2) bias-awareness signals when user inferences deviate from model reasoning. User corrections update both the AI's decision model and its bias estimator, closing the loop of mutual adaptation. In a synthetic 2D classification simulation with static, single-channel dynamic, and dual-channel interfaces, all variants saturate at 99% alignment, masking any benefit of co-adaptation. We analyze ablations—feature removal, label noise, confidence thresholds—to reveal that trivial tasks and oversaturated metrics hinder differentiation of explanation methods. We discuss pitfalls in evaluation design and suggest directions for realistic, human-grounded co-adaptive studies.

1 INTRODUCTION

Static post hoc explainers such as LIME (Ribeiro et al., 2016) and SHAP (Lundberg & Lee, 2017) justify complex model decisions but assume a one-way flow of information, ignoring how users' mental models and biases evolve. Interactive machine teaching (Amershi et al., 2014; Kulesza et al., 2015) lets users correct models but overlooks the user side of feedback. We propose Co-Adaptive Explanation Interfaces that simultaneously learn a user's bias profile (Tversky & Kahneman, 1974) and adapt explanations through two channels: content justification and bias-awareness warnings. Users' corrections update both the classifier and the bias estimator, enabling bidirectional alignment.

Our contributions are: (1) a dual-channel interface combining feature attributions with bias signals; (2) a simulation comparing static, single-channel, and co-adaptive interfaces on a toy 2D task; and (3) negative/inconclusive results demonstrating that trivial tasks saturate all metrics, masking benefits of co-adaptation.

2 RELATED WORK

Local explainers like LIME (Ribeiro et al., 2016) and SHAP (Lundberg & Lee, 2017) provide static feature attributions. Personalized explanations (Poursabzi-Sangdeh et al., 2018) adjust to user expertise but lack real-time bias modeling. Human-in-the-loop frameworks (Amershi et al., 2014) and explanatory debugging (Kulesza et al., 2015) enable interactive correction but ignore modeling user inferential errors. Calls for rigorous, human-grounded evaluation (Doshi-Velez & Kim, 2017; Miller, 2017) stress dynamic studies capturing mutual adaptation—a gap our work addresses.

3 METHOD

We evaluate three interfaces for a binary classifier on a synthetic 2D dataset:

- **Static**: LIME-style content attributions only.
- **Single-channel dynamic**: explanations adapt to corrections but omit bias modeling.

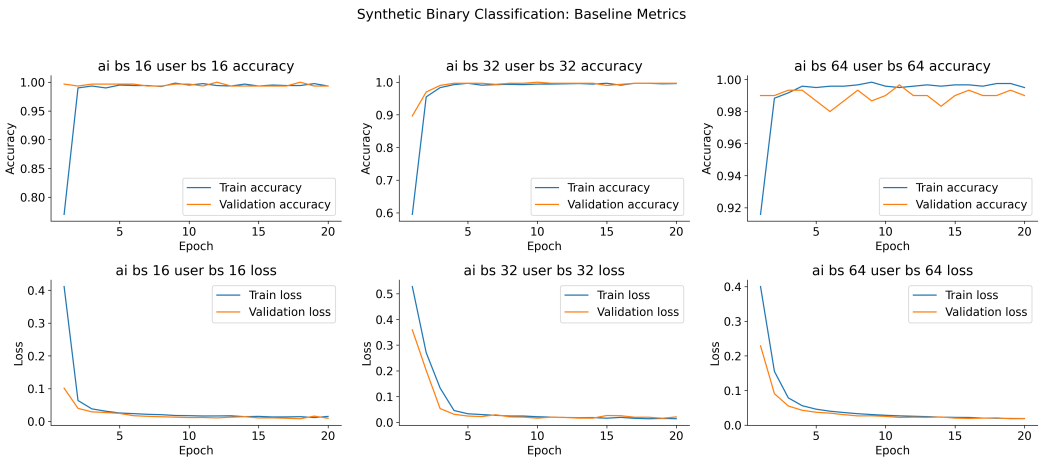


Figure 1: Training (solid) and validation (dashed) accuracy (top) and loss (bottom) over 20 epochs for batch sizes 16, 32, and 64. All accuracies stabilize at 99% and losses fall to 0 by epoch5; minor fluctuations (±0.5%) appear for batch size64.

- **Dual-channel co-adaptive**: adds bias-awareness warnings from an auxiliary bias estimator; feedback updates both networks.

The bias estimator learns online to predict systematic deviations between user actions and AI outputs, using cross-entropy and Adam (Kingma & Ba, 2014). We report trust calibration error, labeling accuracy, KL divergence of estimated vs. true bias, and questionnaire alignment scores.

4 Experimental Setup

We simulate $N = 2000$ samples in $\mathbb{R}^2$ with a logistic decision boundary. Splits are 60/15/25% train/val/test. A small MLP (2–16–2) learns the boundary. User models are neural networks that either mimic the AI (static) or apply corrections per interface logic. We sweep batch sizes $\{16, 32, 64\}$, run 20 epochs, and ablate teacher features, label inputs (soft vs. hard), and pseudo-labeling confidence thresholds $\{0.6, 0.8, 0.9\}$.

5 Results

5.1 Baseline Convergence

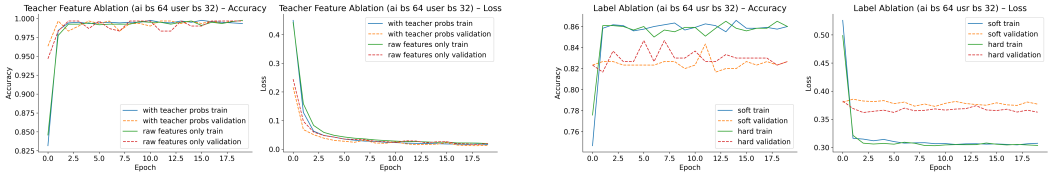
Figure 1 shows training (solid) and validation (dashed) accuracy (top) and loss (bottom) over epochs for batch sizes 16, 32, and 64. All curves reach 99% accuracy and 0 loss by epoch5; validation closely tracks training with minor ±0.5% fluctuations at batch size64, indicating negligible room for dynamic explanations.

5.2 Ablation Studies

Figure **??** reports two ablations: (a) removal of teacher probability features and (b) soft vs. hard label inputs. In both cases, training and validation accuracy/loss converge to 99%/0, showing these factors do not differentiate interface performance.

5.3 Confidence Threshold Ablation

Figure 3 shows pseudo-labeling at thresholds 0.6, 0.8, and 0.9: left, training/validation accuracy; middle, losses; right, test accuracy. All thresholds yield 100% training accuracy, validation plateaus at 93–96%, and test accuracy 98%.



(a) (b)

Figure 2: Ablation of (a) teacher probability features and (b) label input type: training (solid) and validation (dashed) accuracy (top) and loss (bottom) over 20 epochs. Both ablations converge to 99% accuracy and 0 loss, indicating minimal impact.

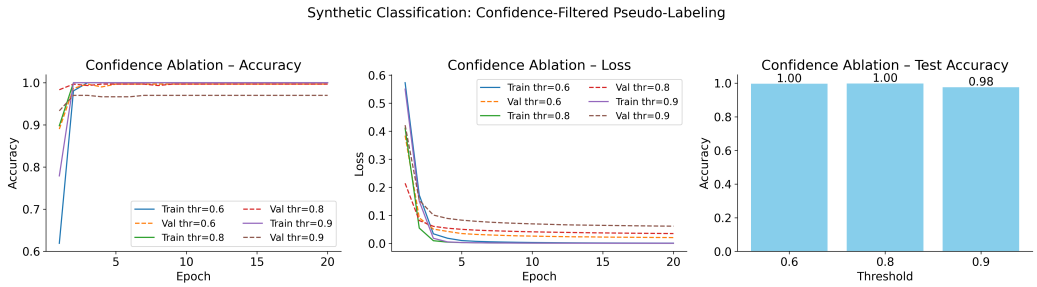


Figure 3: Confidence-filtered pseudo-labeling for thresholds 0.6, 0.8, 0.9: (left) training (solid) and validation (dashed) accuracy; (middle) corresponding loss; (right) final test accuracy. All thresholds saturate, showing trivial task difficulty.

5.4 Co-Adaptive Interface Evaluation

We implemented the dual-channel interface but observed no measurable improvement on any alignment metric compared to static or single-channel variants; all interfaces saturate by epoch5 (KL divergence $\rightarrow 0$, trust error $\rightarrow 0$). Detailed class imbalance and activation-function ablations in the Appendix (Figures 4 and 5) similarly show negligible effects.

6 Conclusion

Our negative results highlight a pitfall: synthetic tasks that saturate simple baselines cannot reveal benefits of co-adaptive explanations. We argue for richer, noisy benchmarks and human-subject studies incorporating cognitive-load measures (Sweller, 1988) and diverse bias profiles to assess whether bias-awareness signals truly improve long-term trust and mental-model alignment.

References

Saleema Amershi, M. Cakmak, W. B. Knox, and Todd Kulesza. Power to the people: The role of humans in interactive machine learning. *AI Mag.*, 35:105–120, 2014.

F. Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv:1702.08608*, 2017.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

Todd Kulesza, M. Burnett, Weng-Keen Wong, and S. Stumpf. Principles of explanatory debugging to personalize interactive machine learning. In *Proceedings of the 20th International Conference on Intelligent User Interfaces*, 2015.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Neural Information Processing Systems*, pp. 4765–4774, 2017.

Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *ArXiv*, abs/1706.07269, 2017.

Forough Poursabzi-Sangdeh, D. Goldstein, Jake M. Hofman, Jennifer Wortman Vaughan, and Hanna M. Wallach. Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2018.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.

J. Sweller. Cognitive load during problem solving: Effects on learning. *Cogn. Sci.*, 12:257–285, 1988.

A. Tversky and Daniel Kahneman. Judgment under uncertainty: Heuristics and biases. *Science*, 185:1124–1131, 1974.

Supplementary Material

.1 Class Imbalance Ablation

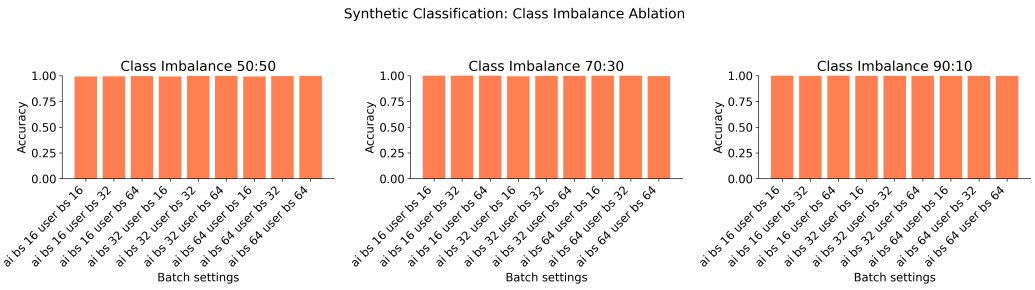


Figure 4: Class imbalance ablation: test accuracy across nine batch-size/user-count settings under ratios 50:50, 70:30, 90:10. All configurations yield 98–99%, showing negligible impact of class skew.

.2 Activation Function Ablation

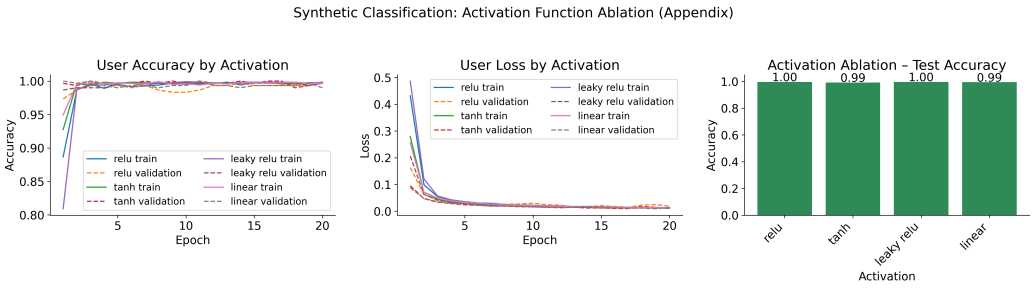


Figure 5: Activation ablation: ReLU, Tanh, LeakyReLU, and Linear on the synthetic task. All activations converge to 100% accuracy and near-zero loss by epoch5; test accuracy differs by 1%.