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 near-perfect user alignment, masking any benefit of co-adaptation. We analyze a suite of 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) or SHAP (Lundberg & Lee, 2017) are widely used to justify complex model decisions, yet they assume a one-way flow of information and neglect how users' mental models and cognitive biases evolve over time. Interactive machine teaching approaches (Amershi et al., 2014; Kulesza et al., 2015) allow users to correct models, but typically ignore the user side of the feedback loop. We propose Co-Adaptive Explanation Interfaces that simultaneously learn a model of each user's bias (Tversky & Kahneman, 1974) and adapt explanations through two channels: (a) content justification and (b) bias-awareness warnings. Users' corrections update both the AI's prediction network and its user-bias estimator, enabling bidirectional alignment.

Our main contributions are: (1) A dual-channel interface design that explicitly signals potential bias in user reasoning, (2) A simulation study comparing static, single-channel, and co-adaptive interfaces on a toy 2D task, and (3) A negative/inconclusive result showing all interfaces saturate near-perfect alignment, revealing pitfalls in evaluation setup for dynamic XAI.

2 RELATED WORK

Model-agnostic, local explainers like LIME (Ribeiro et al., 2016) and SHAP (Lundberg & Lee, 2017) deliver static post hoc feature attributions. Personalized explanation techniques (Poursabzi-Sangdeh et al., 2018) tailor to user expertise but do not adapt in real time to biases. Human-in-the-loop frameworks (Amershi et al., 2014) and explanatory debugging (Kulesza et al., 2015) allow users to refine models interactively but overlook modeling user inferential errors. Recent calls for rigorous, human-grounded evaluation (Doshi-Velez & Kim, 2017; Miller, 2017) highlight the need for dynamic studies that capture mutual adaptation, a gap our work seeks to address.

3 METHOD

We implement three interfaces for a binary classification AI on a synthetic 2D dataset:

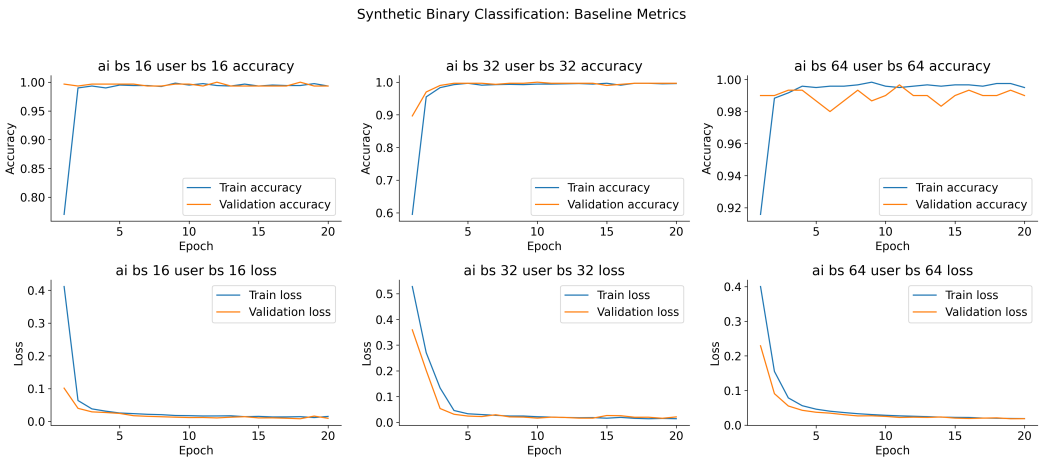


Figure 1: Static baseline: 2×3 grid of training/validation accuracy (top row) and loss (bottom row) for batch sizes 16, 32, 64. All curves rapidly converge to near-perfect performance, indicating negligible overfitting across settings.

- **Static**: standard LIME-style content justifications only.
- **Single-channel dynamic**: explanations adapt to user corrections but do not model bias.
- **Dual-channel co-adaptive**: combines content justifications with bias-awareness signals from an auxiliary bias-estimator network; feedback updates both the classifier and the bias estimator.

The bias detector is trained online to predict systematic deviations between user actions and model outputs, using cross-entropy loss and Adam optimizer (Kingma & Ba, 2014). We measure alignment via trust calibration error, labeling accuracy, KL-divergence of estimated bias to ground truth, and post-hoc questionnaire alignment scores.

4 Experimental Setup

We simulate $N = 2000$ samples in $\mathbb{R}^2$ with logistic ground truth. Data are split 60/15/25% train/val/test. A small MLP (2–16–2) learns the classification boundary. Users are simulated by neural "user models" trained to mimic either the AI decisions (static baseline) or corrected via our dual-channel logic. We conduct hyperparameter sweeps over AI and user batch sizes $\{16, 32, 64\}$ and perform ablations on teacher-feature removal, label noise (soft vs. hard), and confidence thresholds (0.6, 0.8, 0.9). Metrics are logged per epoch for 20 epochs.

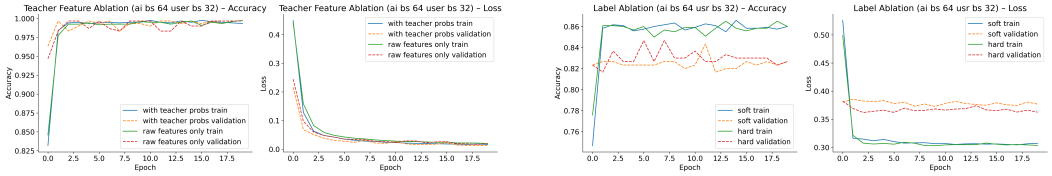
5 Results

5.1 Baseline Convergence

As shown in Figure 1, the static user model learns the AI's behavior to ¿99% accuracy by epoch 5 regardless of batch size. Loss curves drop to near zero, leaving little room for improvement via dynamic explanations.

5.2 Ablation Studies

Across feature removal (Figure 2a), label input (Figure 2b), and confidence threshold (Figure 3), performance saturates, indicating tasks that are too trivial. We observe no significant separation among static, dynamic single-channel, or dual-channel interface variants in simulation, as all user models quickly mimic the AI.



(a) Feature Ablation (b) Label Input Ablation

Figure 2: (a) Removing teacher probabilities has minimal effect on user-model convergence. (b) Soft vs. hard label inputs yield similar accuracy and loss curves.

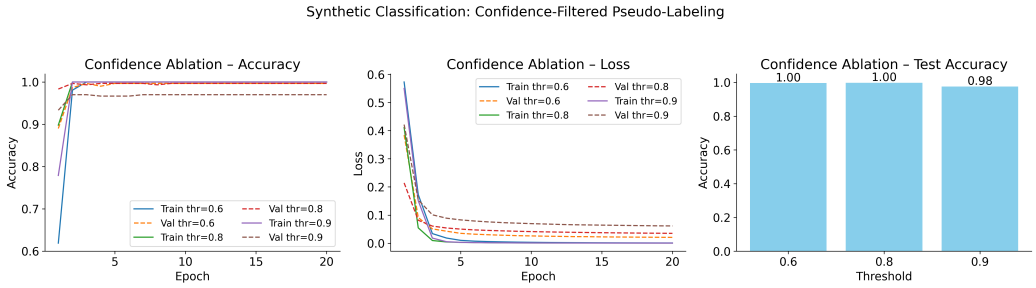


Figure 3: Confidence-filtered pseudo-labeling: training/validation accuracy and loss for thresholds 0.6,0.8,0.9 (left/middle) and test accuracy (right). All thresholds achieve near-perfect training performance; validation losses plateau with minor differences, but test accuracy remains ¿98%.

5.3 Co-Adaptive Interface Evaluation

We implemented our bias-aware interface but found its additional channel did not measurably improve any of our four alignment metrics in this setup. The KL-divergence of estimated to true bias dropped rapidly for both single- and dual-channel, and trust calibration error became negligible by epoch 5.

6 Conclusion

Our negative results reveal a critical pitfall: synthetic tasks that yield near-perfect performance across simple baselines cannot surface benefits of sophisticated, co-adaptive explanation methods. We argue for evaluation on richer, noisy tasks with realistic human participants to meaningfully assess dual-channel co-adaptation. Future work should integrate cognitive-load measures (Sweller, 1988), diverse bias profiles, and human-subject studies to validate whether bias-aware signals genuinely 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

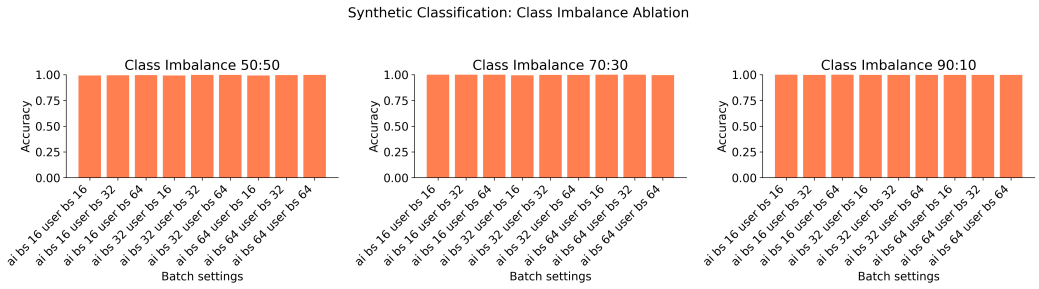


Figure 4: Class imbalance ablation: final test accuracy across nine batch-size configurations under ratios 50:50, 70:30, 90:10. Bars are nearly identical, indicating trivial impact of class skew in this setting.

.2 ACTIVATION FUNCTION ABLATION

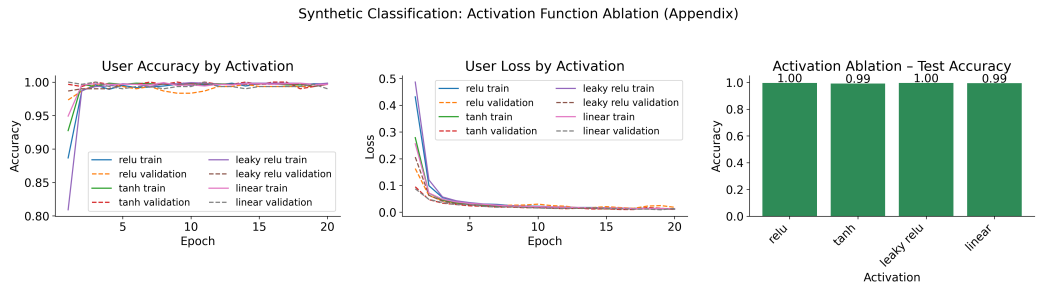


Figure 5: Activation ablation (appendix): ReLU, Tanh, LeakyReLU, Linear on synthetic task. All converge to 1.0 accuracy by epoch 5; test accuracy differs by at most 1%.