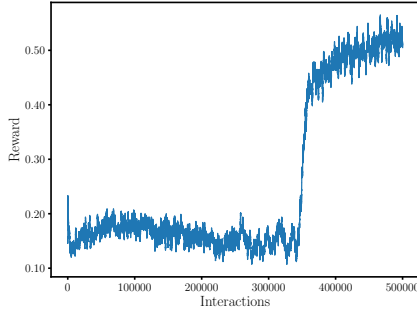


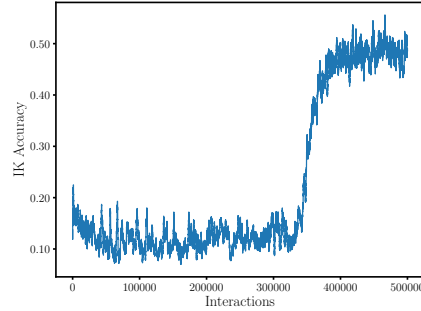
A APPENDIX

A.1 ADDITIONAL RESULTS FOR THE COVERTYPE IGL SIMULATION

Model Parameters For The Covertypes IGL Simulation. For the Covertypes IGL simulation we trained both a CB and IK model (as shown in Algorithm 1). We use online multiclass learning to implement the IK model, where the classes are all the possible actions and the label is the action played. Both CB and IK are linear logistic regression models implemented in PyTorch, trained using the cross-entropy loss. Both models used Adam to update their weights with a learning rate of $2.5e^{-3}$. The first layer of weights for both models had 512 weights that were initialized using a Cauchy distribution with $\sigma = 0.2$. The learning curves are shown in Figure 5b



(a) Reward learning.



(b) Accuracy of the inverse kinematics model used for the intermediate task of action prediction.

Figure 5: Learning curves for the simulated experiment in Sec. 4.1.

A.2 ADDITIONAL RESULTS FOR FACEBOOK NEWS SIMULATION

Model Parameters For The Facebook News Simulation. For the Facebook news simulation there were 2 CB models and an IK model. All models used Adam to update their weights with a learning rate of 10^{-3} , batch size 100, and cross-entropy loss function. The reward learning curves are shown in Figures 6,7 and 8. Further details about implementation are available in provided code.

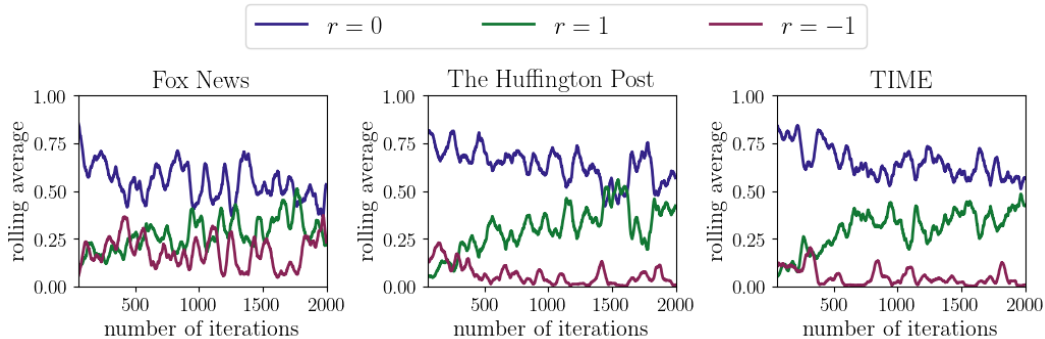


Figure 6: Reward learning curves for IGL-P (3) in Facebook news recommendation experiment. Over time, IGL-P (3) learns the difference between the three states and tries to minimize $r = 0$ and $r = -1$ while maximizing $r = 1$ across all contexts.

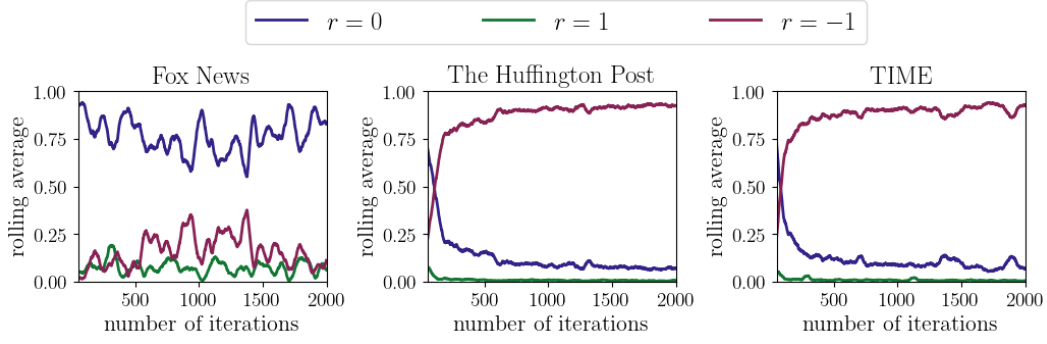


Figure 7: Reward learning curves for CB-emoji in Facebook news recommendation experiment. In two of the three different contexts (The Huffington Post readers and TIME readers), CB-emoji tries to and succeeds in maximizing $r = -1$.

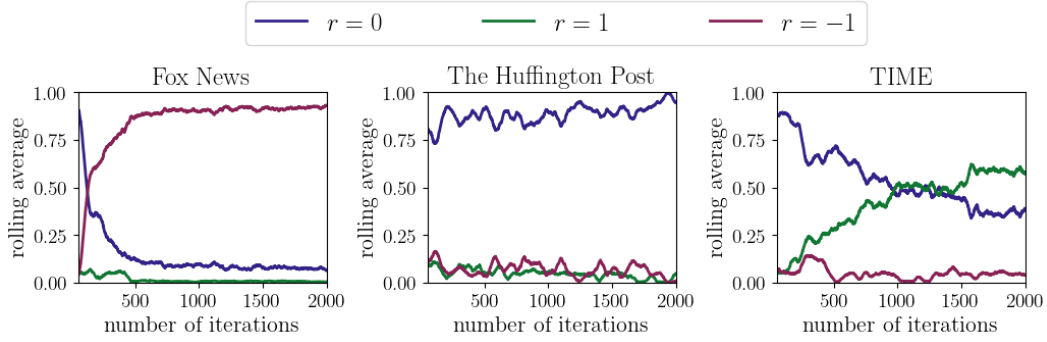


Figure 8: Reward learning curves for CB-comment in Facebook news recommendation experiment. Although CB-comment performs successfully with TIME readers, it maximizes the unhappiness of Fox News readers and promotes indifference in The Huffington Post readers.