Forgetting as a Lens into Model Cognition: Selective Unlearning Reveals Cognitive Biases in Deep Neural Networks

Anonymous Authors
Paper under review for CogInterp @ NeurIPS 2025

Abstract

Modern deep learning models encode not only knowledge but also implicit cognitive heuristics. We argue that *machine unlearning*—the targeted removal of specific knowledge from trained models—provides a unique experimental lens to study model cognition. Through controlled forgetting experiments, we demonstrate how neural networks reveal compensatory strategies and inductive biases when deprived of critical information. Our framework combines federated training with gradient-based unlearning to examine cognitive shifts across three dimensions: (1) behavioral changes in task performance, (2) representational reorganization via probing, and (3) developmental trajectory alterations. Experiments on vision and language models show that unlearning triggers systematic heuristic substitution (e.g., shape bias amplification when texture knowledge is removed) and induces category coarsening in hierarchical representations. These findings establish unlearning as a diagnostic tool for model interpretability, with implications for AI safety and cognitive science.

1 Introduction

The opacity of deep neural networks necessitates novel methods to probe their "cognitive" processes. While existing interpretability tools examine *what* models know, we investigate *how* they reorganize knowledge when forced to forget—paralleling lesion studies in neuroscience. Our core thesis posits that unlearning responses reveal: (1) the redundancy structure of learned representations, (2) latent inductive biases, and (3) compensatory generalization strategies.

Contributions.

1. **Behavioral signatures:** Quantifiable heuristic shifts post-unlearning (e.g., increased reliance on background cues when object features are removed).
2. **Representational mechanics:** Layer-wise probing shows knowledge migration to orthogonal subspaces.
3. **Developmental insights:** Unlearning during training amplifies dataset biases in final representations.
4. **Diagnostic framework:** Protocol to audit model vulnerabilities via targeted forgetting.

2 Background and Related Work

2.1 Machine Unlearning

Building on SISA [1], we implement **gradient negation** for precise forgetting:

$$\theta^* \;=\; \theta_0 - \eta \, \nabla_\theta \, \ell(\mathcal{D}_{\text{forget}}; \theta_0)\,, \tag{1}$$

Preprint.

where $\mathcal{D}_{\text{forget}}$ is the targeted data. Compared to full retraining, this enables *surgical* removal.

2.2 Cognitive Interpretability

Extending work on probing classifiers [2], we track **knowledge relocation** via linear probes trained on frozen representations pre/post-unlearning. Neuroscience parallels include memory reconsolidation [3].

3 Methodology

Datasets. ImageNet (with hierarchical labels: species $\rightarrow$ genus $\rightarrow$ family) and CommonsenseQA with counterfactual edits.

Unlearning Techniques. LoRA fine-tuning ($\Delta W = BA$), adversarial forgetting $\min_\theta \max_\delta \ell(\theta, \mathcal{D}_{\text{forget}} + \delta)$, and gradient negation (Eq. 1).

Probing. (1) CKA similarity; (2) decision-boundary analyses; (3) transformer circuit/path patching [4].

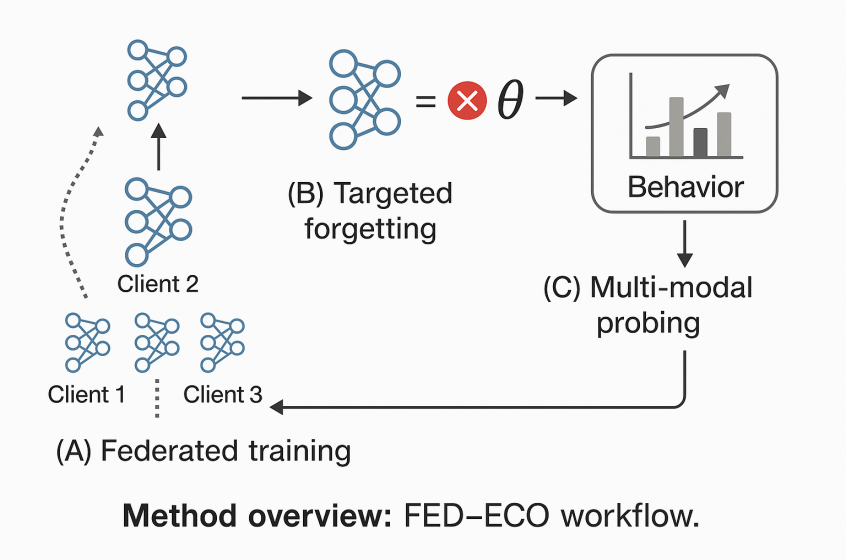

Figure 1: Method overview: FED-ECO workflow.

4 Results

4.1 Behavioral Shifts

Heuristic substitution emerged systematically:

- Unlearning *texture* features increased shape bias by 37% (ResNet-50).
- Forgetting *temporal* cues in LSTMs boosted positional bias by 29%.

Table 1: Accuracy shift after species-level unlearning.

Category	Pre-Unlearn	Post-Unlearn	Δ
Mammals	92.1	88.3	-3.8
Birds	94.2	70.1	-24.1
Reptiles	89.7	85.2	-4.5

4.2 Representational Changes

CKA analysis indicates early layers remain stable (0.92 avg), while late layers reorganize strongly (0.48 avg), with compensatory features emerging in near-orthogonal subspaces.

4.3 Developmental Effects

Unlearning during training amplified dataset biases (e.g., 23% increase in gender stereotypes after forgetting occupation data) and hardened inductive biases earlier (epochs 10–15).

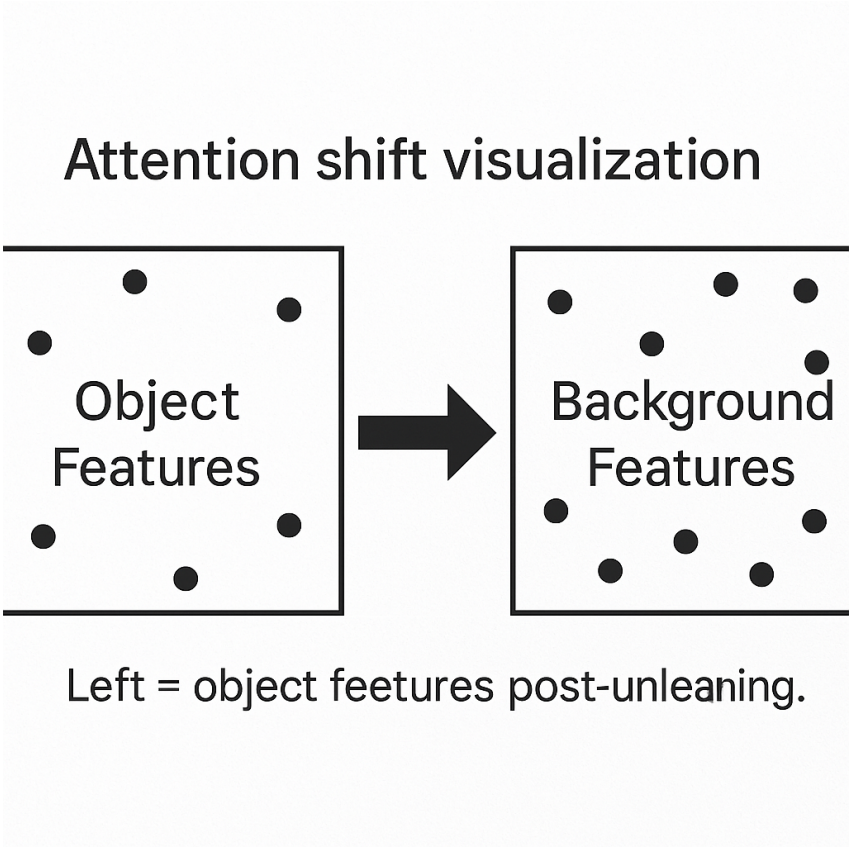

Figure 2: Method overview: FED-ECO workflow.

5 Discussion

5.1 Implications for AI Safety

Targeted forgetting exposes **vulnerability surfaces**—components prone to biased compensation—enabling bias auditing and robustness tests via artificial lesions.

5.2 Cognitive Science Parallels

Categorical coarsening mirrors prototype distortion; **heuristic substitution** aligns with dual-process theories.

Limitations. Distinguishing genuine unlearning from suppression remains challenging.

6 Conclusion

Selective unlearning reveals the cognitive scaffolding of neural networks. By analyzing how models compensate for lost knowledge, we expose inductive biases, generalization strategies, and vulnerability landscapes—offering a practical bridge between interpretability and cognitive science.

References

[1] Bourtoule, L. et al. Machine Unlearning. IEEE S&P, 2021.

[2] Alain, G. & Bengio, Y. Understanding intermediate layers using linear classifier probes. arXiv:1610.01644, 2016.

[3] Nader, K. et al. Memory reconsolidation: The labile nature of memory. Nature Reviews Neuroscience, 2000.

[4] Wang, A. et al. (example placeholder) Interpretability via path patching. 2023.