Diagnosing Shortcut-Induced Rigidity in Continual Learning: The Einstellung Rigidity Index (ERI)

Kai Gu* and Weishi Shi
Department of Computer Science and Engineering
University of North Texas, Denton, TX 76203, USA
Email: {kaigu@my.unt.edu, weishi.shi@unt.edu}

***Abstract*—Deep neural networks frequently exploit shortcuts—incidental correlations such as distinctive colors, textures, or background artifacts that predict labels without causal meaning—which undermines robustness and out-of-distribution reliability. In continual learning (CL), these shortcuts can persist and intensify: weights inherited from earlier tasks bias representation reuse toward whatever features most easily satisfied prior labels, mirroring the cognitive *Einstellung* effect in which past habits block optimal solutions. Whereas catastrophic forgetting erodes past skills, shortcut-induced rigidity throttles the acquisition of new ones. We introduce the Einstellung Rigidity Index (ERI), a compact diagnostic that disentangles genuine transfer from cue-inflated performance using three interpretable facets: (i) Adaptation Delay (AD), (ii) Performance Deficit (PD), and (iii) Relative Suboptimal Feature Reliance (SFR_{rel}). On a two-phase CIFAR-100 CL benchmark with a deliberately spurious magenta patch in Phase 2, ERI flags cases where pretraining accelerates shortcut reuse but impairs semantic learning. Across Naive, EWC, and Replay, we observe high accuracy on patched images but collapse when the patch is masked, large negative AD (faster threshold attainment due to shortcut reuse), negative PD (seeming gains inflated by the cue), and positive SFR_{rel} (compared to a Scratch-T2 baseline). While these findings are consistent with shortcut-induced rigidity, we caution that results may vary with backbones, datasets, and cue salience. ERI provides a conservative litmus test to screen tasks for rigidity risks and guide interventions.**

***Index Terms*—Continual learning; shortcut learning; spurious correlations; Einstellung effect; robustness; evaluation metrics; representation similarity.**

I. Introduction

Deep networks often latch onto easy-to-exploit correlations that are predictive but non-causal, a phenomenon termed *shortcut learning* [1]–[3]. Such dependencies can inflate in-distribution accuracy while degrading robustness and generalization. In continual learning (CL), where models must acquire new tasks sequentially, shortcut use may be further entrenched because previously consolidated representations bias the search toward low-effort features. This dynamic mirrors the cognitive *Einstellung effect*, where prior solution strategies impede discovery of better ones [4].

Contemporary CL evaluation focuses on aggregated metrics such as average accuracy (ACC), backward transfer (BWT), and forward transfer (FWT) [5]. While invaluable, these do not distinguish whether success stems from robust, causal features or from spurious cues [6]–[9]. When spurious correlations are present, they can both accelerate learning (by supplying an easy signal) and impair learning (by suppressing gradients for more informative features). The result is misleading progress indicators and reduced scalability.

We propose the *Einstellung Rigidity Index* (ERI): a three-facet diagnostic designed to expose shortcut-induced rigidity in sequential learning. ERI quantifies: (i) how quickly a continual learner reaches an accuracy threshold relative to a from-scratch baseline on shortcut-bearing data (Adaptation Delay, AD); (ii) the final accuracy gap on those data (Performance Deficit, PD); and (iii) how much more the continual model relies on the shortcut than the from-scratch baseline, measured via a masking intervention (Relative Suboptimal Feature Reliance, SFR_{rel}). We validate ERI on CIFAR-100 with a ResNet-18 backbone [10], [11], injecting a fixed 4×4 magenta patch into select Phase 2 classes. Our analyses suggest ERI can differentiate genuine transfer from shortcut reuse. Still, we avoid over-claiming: results should be reexamined under different architectures, tasks, and naturally occurring artifacts.

Contributions: (1) ERI: a compact, interpretable diagnostic for shortcut-induced rigidity in CL. (2) A controlled two-phase benchmark with masked evaluations to isolate shortcut reliance. (3) Evidence that Naive, EWC [12], and Replay can exhibit negative AD, negative PD, and elevated SFR_{rel}, indicative of rigidity. (4) Practical guidance: screening protocols and mitigation strategies drawing on mutual-information minimization [13], [14] and disentanglement [15], with curriculum suggestions and attention stability considerations [8].

II. Related Work

A. Shortcut learning and spurious correlations

Shortcut learning is widely observed in vision and medical imaging [1]–[3], where models exploit textures, color patches, or device-specific artifacts. Proxies for shortcut detection include mutual information [13], [14] and counterfactual

interventions [15]. Temporarily disabling shortcuts can aid learning [16], though in streams this may be nontrivial.

B. Continual learning and evaluation

CL methods mitigate forgetting via parameter regularization (e.g., EWC [12]), replay, prompting, or modularity [5]. Yet evaluation often conflates semantic mastery with shortcut exploitation. Attention stability [8] and optimizer choice [6] can affect forgetting dynamics, while recent NLP analyses probe how memories are revived [9]. We position ERI as a complementary diagnostic to standard ACC/BWT/FWT.

III. The Einstellung Rigidity Index (ERI)

We formalize ERI for a Phase 2 scenario where certain classes are augmented with a spurious patch. Let $\mathcal{A}_S(e)$ and $\mathcal{A}_{CL}(e)$ denote the test accuracies on patch-bearing Phase 2 data after e epochs for a Scratch-T2 model M_S and a continual model M_{CL}, respectively. Fix a threshold τ.

A. Adaptation Delay (AD)

Define the epochs to reach τ as:

$$E_S(\tau) = \min\{e \mid \mathcal{A}_S(e) \geq \tau\}, \quad E_{CL}(\tau) = \min\{e \mid \mathcal{A}_{CL}(e) \geq \tau\} \tag{1}$$

The AD is

$$\text{AD} = E_{CL}(\tau) - E_S(\tau). \tag{2}$$

A substantially negative AD indicates the continual learner reaches τ faster, which may reflect shortcut reuse rather than semantic transfer when paired with high shortcut reliance.

B. Performance Deficit (PD)

Let $\mathcal{A}_S^*$ and $\mathcal{A}_{CL}^*$ be the final test accuracies (patch present). Define

$$\text{PD} = \mathcal{A}_S^* - \mathcal{A}_{CL}^*. \tag{3}$$

A negative PD means the continual model appears better on patched data. Combined with masking analyses, this can reveal cue-inflated gains.

C. Relative Suboptimal Feature Reliance (SFR_{rel})

For a model M, define

$$\Delta_M = \mathcal{A}_M^{\text{patch}} - \mathcal{A}_M^{\text{masked}}, \tag{4}$$

the performance drop when the shortcut is removed for the designated T2 classes. The relative reliance is

$$\text{SFR}_{\text{rel}} = \Delta_{CL} - \Delta_S. \tag{5}$$

Positive SFR_{rel} indicates the continual process amplified dependence on the shortcut relative to the scratch baseline.

Interpreting ERI jointly: Interpreting the three facets together is essential: - Red-flag pattern: AD $\ll 0$, PD ≤ 0, and $\text{SFR}_{\text{rel}} > 0$ suggest shortcut-accelerated learning and brittle performance. - Benign transfer: AD ≈ 0 or > 0, PD ≥ 0, $\text{SFR}_{\text{rel}} \leq 0$ indicates no elevated shortcut reliance. - Ambiguous cases warrant additional probes (e.g., CKA drift, calibration, counterfactual data).

TABLE I
Mean results across four seeds ($\pm$ SD). SC=T2 shortcut classes; NSC=T2 non-shortcut classes. AD computed only where Scratch-T2 reached $\tau = 0.6$ (n=2).

Strategy	AD	PD	SFR_{rel}	$\mathcal{A}_{\text{SC}}^{\text{norm}}$	$\mathcal{A}_{\text{SC}}^{\text{mask}}$	$\mathcal{A}_{\text{T1}}$	$\mathcal{A}_{\text{NSC}}^{\text{norm}}$
Scratch_T2	N/A	N/A	0.000	0.586	0.000	N/A	0.617
Interleaved	-6.00	-0.089	0.093	0.679	0.000	0.329	0.353
Naive_Seq	-14.0	-0.144	0.142	0.730	0.002	0.000	0.664
EWC_Seq	-14.0	-0.156	0.146	0.742	0.009	0.000	0.674
Replay_Seq	-14.0	-0.156	0.155	0.742	0.001	0.000	0.664

AD: epochs (where $\tau = 0.6$). Values are mean $\pm$ SD across seeds (omitted to conserve space; see Appendix for full values).

IV. Experimental Setup

A. Dataset, tasks, and shortcut injection

We construct a two-phase CL benchmark on CIFAR-100 [10]. Phase 1 (T1) comprises N_1 base superclasses using natural images. Phase 2 (T2) introduces N_2 new superclasses, with five designated "shortcut T2 classes" augmented by a fixed 4×4 magenta square in the top-left of every training and evaluation image. The patch is chosen for high contrast and no semantic meaning.

B. Models and baselines

We use ResNet-18 [11]. We evaluate Naive sequential fine-tuning (Naive_Seq), EWC [12] (EWC_Seq), and Experience Replay (Replay_Seq), and compare against: (i) Scratch-T2 (M_S), trained only on Phase 2; (ii) Interleaved (Joint) training on Phase 1+2, approximating an oracle with unlimited memory. The gap between Interleaved and Scratch-T2 loosely bounds the feasible region for realistic methods.

C. Training, logging, and probes

We track epoch-wise accuracy and loss for patched and masked conditions, store confusion matrices, and compute CKA similarities between early/late layers to assess representational drift. We mask the patch at test time for the designated classes by overwriting the 4×4 region with neutral pixels. We report ERI metrics per seed and aggregate across four seeds.

D. Implementation details

Unless noted, we use standard data augmentation (random crop/flip), SGD with momentum, cosine LR schedule, and cross-entropy loss. Replay buffers use class-balanced sampling when present. EWC hyperparameters are tuned on a held-out split of Phase 2. Hardware and runtimes are logged to support reproducibility.

V. Results

Across four seeds, every sequential strategy attained $\geq 68\%$ accuracy on patched images yet collapsed to $\leq 1\%$ after masking, indicating strong dependence on the patch and poor semantic learning for the shortcut T2 classes [2]. Sequential methods did not retain Phase 1 knowledge, whereas Interleaved preserved nontrivial T1 performance.

Fig. 1. Shortcut injection and masking. Left: example with 4×4 magenta patch (top-left). Right: masked evaluation for the same class. Please replace with high-resolution, English-only figures embedded within the text.

A. ERI facets

Adaptation Delay (AD). AD ranged from -14 to -6 epochs (mean across seeds), indicating faster attainment of the threshold $\tau = 0.6$ by sequential models relative to Scratch-T2. Standing alone this seems advantageous, but combined with masking it suggests shortcut-accelerated learning rather than genuine transfer.

Performance Deficit (PD). PD ranged from -0.16 to -0.09 (better patched accuracy for sequential models). Masking eliminated this advantage, revealing brittleness consistent with shortcut reliance.

$\mathbf{SFR_{rel}}$. Relative reliance spanned 0.093 to 0.155, quantifying how much more sequential training amplified dependence on the patch compared to Scratch-T2.

B. Non-shortcut classes and representational drift

Sequential models exhibited solid accuracy on T2 non-shortcut classes ($\mathcal{A}_{\text{NSC}}^{\text{normal}} \approx 0.66$–$0.67$), indicating capacity for semantic learning in the absence of an overriding shortcut. CKA analysis (not shown) suggested larger drift in later layers during Phase 2 for sequential methods compared to Interleaved, consistent with repurposing higher-level features toward the shortcut.

VI. Discussion

We view pretraining through a dual lens: it can accelerate learning by reusing easy cues, and impair learning by constraining representation updates around prior, non-causal attractors—an *Einstellung-like* rigidity. ERI's joint signals (AD $\ll 0$, PD ≤ 0, $\text{SFR}_{\text{rel}} > 0$) provide a conservative indication that the sequential learner is exploiting the spurious patch rather than acquiring robust semantics.

Why would classical CL methods fare poorly here? - **Naive fine-tuning** has no guardrail; gradients gravitate to the high-signal patch, suppressing signals for useful features. - **EWC** regularizes movement on parameters deemed important for T1, but the Fisher may reflect spurious T1/T2 correlations; when the patch is highly predictive, EWC can still permit bias toward it, while limiting adjustments that would discover alternative features [12]. - **Replay** can reduce forgetting, but unless replay data dilute or counteract the shortcut, it does not prevent shortcut adoption.

Beyond our synthetic patch, shortcuts can be subtler (scanner markers, backgrounds, text overlays) [3]. Simple accuracy gains should be treated cautiously unless supported by counterfactual tests, calibration checks, or feature-attribution sanity checks [2]. Mutual-information monitoring can flag high cue–label dependence [13], [14], and temporarily disabling shortcuts or randomizing their location can force feature diversification [16]. Attention stability [8] and optimizer choices [6] may modulate rigidity: highly adaptive optimizers can quickly overfit to shortcuts without constraints.

Actionable guidance: - **Screen with ERI.** Before deployment or long training runs, compute ERI on held-out, shortcut-bearing data with masking. Treat the red-flag pattern as a sign to intervene. - **Pair objectives with cue suppression.** Penalize MI between representations and suspected cues [13], [14]; add invariance by data augmentations or counterfactuals; consider disentanglement-based regularization [15]. - **Curriculum design.** Delay exposure to high-salience artifacts, or inject counterfactual samples early to reduce the MI spike. Randomize patch location/appearance during training to prevent hard-coding. - **Evaluation hygiene.** Always report masked performance for suspected shortcut classes, CKA drift, and per-class confusion.

We do not claim ERI replaces standard CL metrics; rather, it complements them by testing for shortcut-induced rigidity. Its utility depends on the quality of masking interventions and the ecological validity of the cue.

VII. Threats to Validity and Limitations

Our cue is synthetic and high-contrast, which may overstate effects compared to subtle, naturally occurring artifacts. We used one backbone (ResNet-18), modest task sizes, and a single patch location; results may change with larger models, patch salience, randomized locations, or domains beyond vision (e.g., language streams). Masking quality matters: imperfect masks could underestimate reliance. We report four seeds but larger runs would improve confidence. We encourage replication on streams such as CORe50 and NICO++ and on textual benchmarks [9].

VIII. Reproducibility Details

We will release code, seed lists, hyperparameters, and data splits. We log training curves, confusion matrices, and CKA statistics per epoch. Patch injection, masking, and ERI computation scripts are modular to ease reuse. All figures should be embedded at high resolution and free of non-English characters.

IX. Conclusion

ERI offers a compact diagnostic to separate genuine transfer from cue-inflated performance in continual learning. In our experiments, ERI flags shortcut-accelerated but brittle learning across several CL strategies. While we are cautious about generalization beyond our setup, we find ERI practical for screening tasks, guiding mitigations, and improving the interpretability of CL evaluations.

Acknowledgment

We thank Weishi Shi, Abdullah Al Forhad, and the Texas Academy of Mathematics and Science for their support.

APPENDIX A
FORMAL ERI DEFINITIONS AND ADDITIONAL RESULTS

We restate ERI definitions with consistent notation and provide per-seed metrics, confusion matrices, and CKA plots in the supplementary materials. Statistical summaries include median and bootstrap 95% CIs for AD, PD, and $\mathrm{SFR}_{\mathrm{rel}}$.

REFERENCES

[1] R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. S. Zemel, W. Brendel, M. Bethge, and F. Wichmann, "Shortcut learning in deep neural networks," *Nature Machine Intelligence*, vol. 2, pp. 665 – 673, 2020. [Online]. Available: https://doi.org/10.1038/s42256-020-00257-z

[2] K. Hermann, H. Mobahi, T. Fel, and M. C. Mozer, "On the foundations of shortcut learning," *ArXiv*, vol. abs/2310.16228, 2023. [Online]. Available: https://arxiv.org/pdf/2310.16228

[3] B. G. Hill, F. L. Koback, and P. L. Schilling, "The risk of shortcutting in deep learning algorithms for medical imaging research," *Scientific Reports*, vol. 14, 2024. [Online]. Available: https://doi.org/10.1038/s41598-024-79838-6

[4] M. Binz and E. Schulz, "Reconstructing the einstellung effect," *Computational Brain Behavior*, vol. 6, pp. 526–542, 2021. [Online]. Available: https://doi.org/10.1007/s42113-022-00161-2

[5] G. M. van de Ven, N. Soures, and D. Kudithipudi, "Continual learning and catastrophic forgetting," *ArXiv*, vol. abs/2403.05175, 2024. [Online]. Available: https://arxiv.org/pdf/2403.05175

[6] D. R. Ashley, S. Ghiassian, and R. S. Sutton, "Does the adam optimizer exacerbate catastrophic forgetting?" 2021. [Online]. Available: https://arxiv.org/pdf/2102.07686

[7] E. Marconato, G. Bontempo, S. Teso, E. Ficarra, S. Calderara, and A. Passerini, "Catastrophic forgetting in continual concept bottleneck models," in *ICIAP Workshops*, 2022. [Online]. Available: https://doi.org/10.1007/978-3-031-13324-4_46

[8] H. Wang, Q. Wu, H. Li, and F. Meng, "Where to forget: A new attention stability metric for continual learning evaluation," in *International Forum on Digital TV and Wireless Multimedia Communication*, 2023. [Online]. Available: https://doi.org/10.1007/978-981-97-3623-2_10

[9] H. Sun and Y. Gao, "Reviving dormant memories: Investigating catastrophic forgetting in language models through rationale-guidance difficulty," *ArXiv*, vol. abs/2411.11932, 2024. [Online]. Available: https://arxiv.org/pdf/2411.11932

[10] A. Krizhevsky, "Learning multiple layers of features from tiny images," 2009. [Online]. Available: https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf

[11] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015. [Online]. Available: http://arxiv.org/abs/1512.03385

[12] J. Kirkpatrick, R. Pascanu, N. C. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, D. Hassabis, C. Clopath, D. Kumaran, and R. Hadsell, "Overcoming catastrophic forgetting in neural networks," *Proceedings of the National Academy of Sciences*, vol. 114, pp. 3521 – 3526, 2016. [Online]. Available: https://doi.org/10.1073/pnas.1611835114

[13] M. Adnan, Y. Ioannou, C.-Y. Tsai, A. Galloway, H. R. Tizhoosh, and G. W. Taylor, "Monitoring shortcut learning using mutual information," *ArXiv*, vol. abs/2206.13034, 2022. [Online]. Available: https://arxiv.org/pdf/2206.13034

[14] L. Fay, E. Cobos, B. Yang, S. Gatidis, and T. Kuestner, "Avoiding shortcut-learning by mutual information minimization in deep learning-based image processing," *IEEE Access*, vol. 11, pp. 64 070–64 086, 2023. [Online]. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=10162210

[15] C. G. Robinson, A. Trivedi, M. Blazes, A. Ortiz, J. Desbiens, S. Gupta, R. Dodhia, P. K. Bhatraju, W. C. Liles, A. Lee, J. Kalpathy-Cramer, and J. M. L. Ferres, "Deep learning models for covid-19 chest x-ray classification: Preventing shortcut learning using feature disentanglement," *PLoS ONE*, vol. 17, 2022. [Online]. Available: https://doi.org/10.1101/2021.02.11.20196766

[16] H. Niu, H. Li, F. Zhao, and B. Li, "Roadblocks for temporarily disabling shortcuts and learning new knowledge," in *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., vol. 35. Curran Associates, Inc., 2022, pp. 29 064–29 075. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2022/file/baaa7b5b5bbaadca5023e1ab909b8af5-Paper-Conference.pdf