Causal Debiasing for Robust Machine Learning

Gaurav Sood*

July 19, 2025

1 Motivation and Background

Modern machine-learning models often appear to generalize well when measured by held-out accuracy, but they can still fail unexpectedly under distribution shifts, adversarial perturbations or fairness criteria. Ribeiro et al. (2020) observed that traditional evaluation overestimates performance because train/validation/test splits share similar biases. Inspired by behavioral testing from software engineering, they proposed *CheckList*, a task-agnostic evaluation matrix consisting of linguistic *capabilities* and *test types* (Ribeiro et al., 2020). Capabilities (e.g., vocabulary, negation, named–entity recognition) correspond to phenomena that a model should handle, while test types check how predictions behave under specific perturbations. The three core test types are:

- **Minimum-Functionality tests (MFTs).** Simple examples designed to probe a specific behavior; they detect when a model uses shortcuts instead of mastering the capability (Ribeiro et al., 2020).
- **Invariance tests (INV).** Label-preserving perturbations that should not change the model's prediction (Ribeiro et al., 2020).
- **Directional Expectation tests (DIR).** Perturbations that should change the label in a known direction, e.g., appending "You are lame" to a positive tweet should decrease the sentiment score (Ribeiro et al., 2020).

By decoupling evaluation from implementation, CheckList reveals failure modes overlooked by aggregate metrics (Ribeiro et al., 2020). However, CheckList focuses on *testing* rather than *training*. This paper explores whether behavioral and causal assumptions can be incorporated into the loss function and training data to improve robustness and fairness, and whether such ideas generalize beyond NLP.

*Gaurav can be reached at gsood07@gmail.com

From a causal-inference perspective, spurious associations arise because the model learns from correlations rather than causal relations. *Negative controls* are well-established tools in epidemiology: researchers design exposures or outcomes that cannot plausibly be causally related to the variable of interest but share the same confounders. Observing an association on a negative control suggests residual confounding (Lipsitch et al., 2010). In experimental biology, negative controls include "inert substance" experiments where the active ingredient is left out; any observed effect under these conditions implies a spurious mechanism (Lipsitch et al., 2010). These ideas motivate adding *falsification tests* to machine–learning pipelines—introducing inputs that should not affect the prediction and penalizing the model when they do.

2 A Unified Loss Framework

Consider a supervised learning task with input x and target y. A trained model $f(x)$ maps inputs to predictions (probabilities for classes). Extending the behavioral and causal considerations above leads to a composite loss:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{task}} + \lambda_{\text{inv}}\,\mathcal{L}_{\text{inv}} + \lambda_{\text{dir}}\,\mathcal{L}_{\text{dir}} + \lambda_{\text{fals}}\,\mathcal{L}_{\text{fals}}. \tag{1}$$

- **Task loss ($\mathcal{L}_{\text{task}}$).** The standard cross–entropy or mean–squared error encourages correct predictions.

- **Invariance penalty ($\mathcal{L}_{\text{inv}}$).** For a label–preserving perturbation $\tilde{x}$ (analogous to CheckList's INV test), penalize the absolute difference in the model's outputs:

$$\mathcal{L}_{\text{inv}} = \mathbb{E}_x\big[\big|f(x) - f(\tilde{x})\big|\big]. \tag{2}$$

  This encourages the model to be insensitive to spurious changes such as swapping gendered pronouns. Garg et al. (2019) use a similar penalty to promote robustness to identity-terms.

- **Directional penalty ($\mathcal{L}_{\text{dir}}$).** For a perturbation x' that is expected to alter the target in a known direction (DIR tests), use a margin–ranking loss:

$$\mathcal{L}_{\text{dir}} = \mathbb{E}_x\,\ell_{\text{ranking}}\big(f(x), f(x')\big), \tag{3}$$

  where $\ell_{\text{ranking}}(p, q) = \max\{0, p - q + \gamma\}$ if the perturbation should lower the score, and $\max\{0, q - p + \gamma\}$ if it should raise it. This enforces monotonic behavior when sentiment is flipped or additional evidence is provided (Ribeiro et al., 2020).

- **Falsification (negative-control) penalty ($\mathcal{L}_{\text{fals}}$).** Inspired by epidemiological negative controls, design *negative-control inputs* x^{neg} that share confounders but should not influence the output. Examples include adding irrelevant words ("table," random

numbers) to text or inserting backgrounds in images that are unrelated to the task. The penalty

$$\mathcal{L}_{\text{fals}} = \mathbb{E}_x\big[\big|f(x) - f(x^{\text{neg}})\big|\big] \tag{4}$$

discourages the model from using these cues; it generalizes the INV penalty to include variables expected to have zero causal effect (Lipsitch et al., 2010).

2.1 Data Augmentation and Synthetic Interventions

Implementing these losses requires mapping inputs to their perturbed counterparts. Two principles guide this process:

1. **Label–preserving perturbations.** These reflect invariances. In NLP, examples include swapping synonyms, changing pronouns, altering named entities, or adding typos. In computer vision, this could involve random cropping, brightness adjustments, or style transfers that preserve the object. However, invariance is only approximate—balancing pronouns may remove gender bias but does not address other confounders.

2. **Causal perturbations.** These represent interventions on the true causal features. In sentiment analysis, flipping adjectives from "amazing" to "terrible" or negating verbs flips the label. In CV, identifying causal features is harder: is a dog's shape causal or its color? Studies show that standard CNNs are strongly biased toward textures rather than shapes (Geirhos et al., 2019). Creating stylized training images that alter textures while preserving shape can increase shape bias and robustness (Geirhos et al., 2019). Synthetic interventions might involve style transfer to change textures, or generating 3D models to change viewpoint.

3. **Negative controls.** These require variables that share the same confounding structure but have no causal relationship. In NLP, one can append random neutral clauses ("by the way, the sky is blue") or numeric strings. In CV, backgrounds, lighting or random overlays can act as negative controls; if the model's prediction changes significantly, it is exploiting spurious cues. Designing effective negative controls demands domain knowledge to ensure they are truly unrelated to the task.

2.2 Extended Falsification and Impossibility Checks

Negative controls typically intervene on the *inputs*—they modify the text or image in ways that should not change the true label. A complementary strategy is to introduce auxiliary tasks or edits that should not be causally affected by the original input and to penalize any association. Two examples illustrate this idea:

Negative-control outcomes. In causal inference, a negative–control outcome is a variable that cannot plausibly be caused by the exposure but may share the same unobserved confounders(Lipsitch et al., 2010). In our setting, we can simulate such outcomes by pairing each training example with a proxy label that is independent of the input. For instance, alongside the sentiment of a movie review, we might ask the model to predict the sentiment of an unrelated filler sentence or the parity of a random number appended to the input. Because there is no causal relationship between the review and this auxiliary target, the expected association is zero. If the model's predictions on the negative–control outcome correlate with the true label, this indicates reliance on spurious features. One can therefore add a penalty term

$$\mathcal{L}_{\text{nc\_out}} = \text{Corr}\big(f_{\text{aux}}(x), y\big)^2 \tag{5}$$

that measures the squared correlation between the auxiliary predictions $f_{\text{aux}}(x)$ and the primary target y, and encourage it to be small. When implementing this idea, care must be taken to ensure the auxiliary task shares the same confounders as the primary task—otherwise the test loses its diagnostic power.

Impossible-effect checks. Whereas directional tests expect monotonic changes in prediction, *impossible effect* checks probe the model's confidence on logically inconsistent or self–contradictory inputs. Inspired by falsification tests in experiments, these examples combine cues that ought to cancel each other out. For instance, the sentence "This movie was wonderful but I hated it" mixes a strongly positive adjective with an explicit negation. A well–behaved sentiment classifier should either output a neutral prediction or at least exhibit uncertainty, not assign extremely high or low probabilities. Similarly, in vision, one might overlay an object with a contradictory label (e.g., placing a dog icon on a picture of a cat) and expect the classifier's confidence to drop. Such examples can be generated systematically to cover different degrees of contradiction. A simple way to enforce appropriate behavior is to penalize high confidence on these inputs with a hinge loss on the predicted logit magnitude:

$$\mathcal{L}_{\text{imposs}} = \mathbb{E}_{x^{\text{imp}}}\big[\max\{0, |f(x^{\text{imp}})| - \tau\}\big], \tag{6}$$

where x^{imp} denotes an impossible–effect example and τ is a confidence threshold. This term discourages the model from making overly confident predictions when the evidence is contradictory. Designing realistic contradictions without introducing unnatural artifacts requires domain knowledge; nonetheless, such checks offer a principled way to probe whether the model respects logical constraints. Both $\mathcal{L}_{\text{nc\_out}}$ and $\mathcal{L}_{\text{imposs}}$ can be incorporated into the overall objective with their own weights (e.g., $\lambda_{\text{nc\_out}}$ and λ_{imposs}) when balancing robustness against the primary task performance.

3 Theoretical Underpinnings and Limitations

3.1 Causal Identification Assumptions

The above framework implicitly assumes we know which features are *causal* for the task. In practice, this may not hold:

- **Ambiguity of causal features.** In images, is the presence of hands in a dog picture a confounder or part of the causal mechanism? The computer–vision literature notes that models may latch onto co–occurring elements (human hands, backgrounds, textures) as confounders (Wang et al., 2021). Without manual annotation or causal models, it is difficult to determine whether shape or color is the relevant causal feature.

- **Partial coverage of confounders.** Balancing pronoun usage (e.g., equal numbers of sentences with "he" and "she") reduces gender bias but does not control for other correlated attributes such as occupation or socio-economic terms. Synthetic data may inadvertently introduce artifacts that the model learns instead.

- **Negative-control validity.** For negative controls to detect confounding, they must share the same confounders as the primary exposure but not causally affect the outcome (Lipsitch et al., 2010). Constructing such variables in high–dimensional domains like images is challenging; backgrounds may not share all confounders, and random overlays could introduce new ones. Falsification penalties therefore provide at most partial assurance.

- **Model capacity and over-regularization.** Adding multiple penalties may degrade performance if hyperparameters are not carefully tuned. A high falsification weight can suppress reliance on legitimate but correlated features.

- **Generalization beyond NLP.** While invariance and directional penalties readily map to textual tasks, designing meaningful interventions in CV or speech requires domain–specific techniques (e.g., style transfer, 3D rendering, audio pitch shifting). The absence of clear linguistic structure makes it harder to interpret and control confounders.

3.2 Connections to Existing Work

This framework unifies several prior ideas:

- **Robustness and adversarial training.** Many robustness techniques augment training with perturbed inputs, penalizing changes in predictions. Counterfactual logit pairing and invariant risk minimization are specific instances of the invariance penalty (Garg et al., 2019).

- **Learning to rank and margin losses.** The directional loss uses a margin–ranking formulation common in information retrieval (Herbrich et al., 2000). CheckList's DIR tests motivate this by treating certain perturbations as "better" or "worse" evidence (Ribeiro et al., 2020).

- **Causal debiasing in CV.** Causal methods for CV often attempt to separate causal features (object shape, salient regions) from spurious correlations (background, texture). Techniques such as the Causal Attention Module (CaaM) model confounders in an unsupervised manner to improve robustness (Wang et al., 2021). Interventional few–shot learning removes confounding from pretrained representations by backdoor adjustment; such methods illustrate how causal reasoning can improve visual recognition (Wang et al., 2021).

- **Negative controls in causal inference.** The falsification penalty echoes epidemiological practice of repeating experiments under conditions expected to yield a null result (Lipsitch et al., 2010). Observing a difference signals confounding or measurement error.

4 Illustrative Examples

4.1 Sentiment Classification

Suppose we train a sentiment classifier. For each training sentence x with label y:

- **Label–preserving perturbations:** swap "he" with "she," replace location names, introduce typos. Compute $\mathcal{L}_{\text{inv}}$ by comparing $f(x)$ and $f(\tilde{x})$ to encourage invariance.

- **Directional perturbations:** flip sentiment–laden adjectives (e.g., "wonderful" $\rightarrow$ "terrible") or add clauses ("I thought it was bad"). Use $\mathcal{L}_{\text{dir}}$ to enforce monotonic decreases or increases as per DIR tests (Ribeiro et al., 2020).

- **Negative controls:** append irrelevant phrases ("by the way, I saw a table") or random numbers that share syntactic structure but no sentiment. Penalize deviations between $f(x)$ and $f(x^{\text{neg}})$.

This combination trains the model to be robust to spurious correlates, sensitive to causal changes, and insensitive to irrelevant information.

4.2 Computer Vision

Consider a dog–breed classifier. Potential interventions include:

- **Label-preserving perturbations:** random cropping, color jitter, background replacement, or stylized images that preserve the dog's shape but alter texture. These correspond to invariance tests.

- **Directional perturbations:** flipping between different breeds is harder, but one could modify the dog's ear shape or fur length in a simulated environment to create examples of another breed; the classifier's confidence should decrease.

- **Negative controls:** add random patches or overlay unrelated objects (e.g., a small colored square) that should not affect the breed. If the classifier's prediction changes, it signals reliance on spurious textures or locations. Researchers have shown that ImageNet–trained CNNs are biased toward recognizing textures rather than shapes (Geirhos et al., 2019); training on stylized images increases shape bias and robustness (Geirhos et al., 2019).

5 Conclusion

Bringing together behavioral testing, causal inference and negative–control principles offers a promising framework for building more robust and fair machine–learning models. *CheckList* highlights the need for comprehensive unit tests across capabilities and perturbations (Ribeiro et al., 2020). Causal inference reminds us to distinguish genuine causal features from spurious correlations and to use negative controls to detect confounding (Lipsitch et al., 2010). The proposed composite loss integrates these insights by penalizing deviations under label–preserving perturbations, enforcing directional behavior when causal features change, and adding falsification penalties to discourage reliance on irrelevant cues.

However, this approach rests on assumptions that are often difficult to satisfy: identifying causal features, designing realistic perturbations, ensuring negative controls share confounders without introducing new ones, and tuning multiple hyperparameters. In high–dimensional domains like images or audio, constructing meaningful interventions requires sophisticated generative models and domain expertise. Thus, while this framework unifies existing ideas and suggests directions for improving robustness and fairness, it should be applied with caution and complemented by careful domain–specific analysis.

References

Garg, S., Perot, V., Limtiaco, N., Taly, A., Chi, E. H., and Beutel, A. (2019). Counterfactual fairness in text classification through robustness. In *Proceedings of the 2019 AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society (AIES)*, pages 219–226.

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., and Brendel, W. (2019). Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations (ICLR)*.

Herbrich, R., Graepel, T., and Obermayer, K. (2000). Large margin rank boundaries for ordinal regression. In *Advances in Large Margin Classifiers*, pages 115–132.

Lipsitch, M., Tchetgen, E. J. T., and Cohen, T. (2010). Negative controls: A tool for detecting confounding and bias in observational studies. *Epidemiology*, 21(3):383–388.

Ribeiro, M. T., Wu, T., Guestrin, C., and Singh, S. (2020). Beyond accuracy: Behavioral testing of nlp models with checklist. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4902–4912.

Wang, T., Xiao, H., Chen, Y., Wei, Y., and Zhang, X. (2021). Causal attention for unbiased visual recognition. *arXiv preprint arXiv:2108.08782*.