Proceedings of the 8th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/2026.36844

Counterfactual Design Fiction Based Identification of Latent Value Biases in Generative Artificial Intelligence

Jinxiu Jiang

Department of Design, Winchester School of Art, University of Southampton, Winchester, UK
jiangjinxiu02530@outlook.com

Abstract. Generative artificial intelligence can produce apparently neutral narratives while embedding unequal assumptions about competence, authority, vulnerability, risk, and institutional support. This study develops a counterfactual design fiction framework for identifying such latent value biases through controlled narrative comparison. A corpus of 1,200 prompt pairs covers employment, healthcare, education, urban services, consumer finance, and domestic technology, with each pair differing in only one social attribute. Llama 3.1 8B Instruct, Qwen2.5 7B Instruct, and Mistral 7B Instruct v0.3 generate 21,600 narratives. Human annotation evaluates agency, competence, risk, resource access, emotional framing, and institutional treatment. Pairwise semantic, role, sentiment, modality, and institutional-action features are integrated through a multi-task Value Bias Identification Network. The proposed model achieved a macro-F1 of 0.882 ± 0.009 and an AUROC of 0.941 ± 0.006, outperforming cosine similarity, logistic regression, XGBoost, and a RoBERTa pair classifier. Socioeconomic status and disability produced the strongest aggregate value disparities, particularly for resource access and institutional treatment. Cross-model experiments retained a macro-F1 above 0.82, indicating that counterfactual comparison captures recurring value structures rather than model-specific lexical artifacts.

Keywords: generative artificial intelligence, design fiction, counterfactual analysis, value bias, computational ethics

1. Introduction

Generative language models are increasingly producing stories, recommendations, explanations, and futures scenarios in socially consequential settings. Linguistic fluency is not necessarily normative neutrality. It has been shown that bias in large language models can manifest itself on the levels of embeddings, probabilities, and generated texts and reproduce unequal association between demographic groups [1]. Social-bias probing additionally shows that there can be systematic differences in model outputs when an identity attribute is varied in the prompt, but other contexts are left unchanged [2].

An especially significant type of bias that is, however, relatively understudied is related to the differential distribution of agency, competence, responsibility, and institutional power. The language-agency analysis has revealed that demographic groups can be represented differently in the roles of leaders, initiators, and supporters [3]. In practice, such biases have also been detected.

DOI: 10.54254/2755-2721/2026.36844

Hiring-related prompts can yield different results depending on the name, associated with race, ethnicity, or gender [4]. Healthcare models have revealed demographic fairness differences for various clinical tasks [5]. There are also geographic and socioeconomic differences in educational recommendation systems [6].

While the majority of existing benchmarks focus on stereotypes, sentiments, toxicity, or recommendations, they do not contain much information regarding whether changes in a social attribute affect narrative causality, tolerated failure, resource availability, blamedness, protection, or decision-making authority. Counterfactual generation provides an opportunity to separate these effects by varying one focal attribute but leaving the surrounding scenario intact [7]. Additionally, design-fiction approach implies putting technology into a context, where users, institutions, conflict, and consequences can be found [8]. Finally, ethics-by-design research implies transformation of abstract values into system behavior [9]. Generated-text studies have shown that bias can occur due to the open-ended narrative structure instead of being based on the classification of the generated text [10]. Thus, the current study relies on the design fiction approach with a pair of counterfactuals.

2. Literature review

The evaluation of biases of generative AI has shifted from lexical associations to generated-text and disparate treatment analyses [1, 2]. This shift is significant since unequal value assignment may be evident in narrative roles and not stereotyping. There may be variations in agency depending on demographic depiction [3], and research on hiring, healthcare, and education reveals that small demographic variations impact decisions and recommendations made by generators [4-6].

The use of counterfactual comparison helps identify these disparities due to its ability to preserve the wider context while varying one social characteristic. However, the created counterfactuals may generate unintended semantic alterations, necessitating the use of strict templates for control [7]. An additional way to evaluate the consequences is design fiction that represents ethical implications through user, technology, institution, and future society interaction [8]. Ethical-by-design frameworks contribute to the operationalization of values into measurable variables [9].

Currently, there are numerous generated-text bias evaluations that concentrate mainly on sentiments, stereotypes, occupations, and decisions [10]. The mentioned characteristics do not include who will gain authority, safety, opportunities, blame, or institutional endorsement. Therefore, the current research framework includes counterfactual comparison with six value dimensions and computational features describing semantic differences, role changes, affective shifts, modalities, and institutions' actions.

3. Method

3.1. Counterfactual design fiction corpus

The corpus contains 1,200 prompt pairs evenly distributed across employment, healthcare, education, urban services, consumer finance, and domestic technology. Each prompt defines a future technology, user objective, institutional setting, decision conflict, and approximately 250-word narrative. Counterfactual variables include gender, age, disability status, socioeconomic position, nationality, and family role, with only one attribute changed within each pair.

Templates are stored in JSONL format. Named entities, occupation, technological function, temporal setting, and requested outcome remain identical across paired prompts. A Python 3.12 checker verifies lexical equivalence outside the counterfactual span and rejects unintended

DOI: 10.54254/2755-2721/2026.36844

substitutions. Fifty pilot pairs are manually reviewed before full generation, ensuring that subsequent narrative differences can be attributed primarily to the manipulated social attribute.

3.2. Controlled narrative generation

Llama 3.1 8B Instruct, Qwen2.5 7B Instruct, and Mistral 7B Instruct v0.3 are executed locally through vLLM 0.6. Each model generates three narratives for both members of every pair, resulting in 21,600 outputs. Sampling uses temperature 0.7, top-p 0.9, a 420-token maximum, and repetition penalty 1.05. System instructions require third-person prose, one identifiable decision event, an institutional response, and a clear consequence for the principal actor. Generation order is randomized independently for model, domain, and counterfactual direction. Model identifiers are removed before annotation. Narratives shorter than 180 words, longer than 320 words, or lacking a detectable decision event are regenerated once. Pair integrity is additionally assessed through named-entity overlap and technology-description similarity so that demographic manipulation does not unintentionally modify the underlying design-fiction scenario.

3.3. Value annotation scheme

There are six dimensions for measuring the latent value allocation that include agency, competence, risk, resource access, emotional framing, and institutional handling. Dimensions are scored on a range from −2 to 2, with negative scores representing a relative disadvantage, 0 representing neutral perception, and positive scores representing an advantage. The other five actors such as the main actor, decision-maker, beneficiary, blameworthy actor, and protected actor are also marked by the annotator. There are three postgraduate annotators who annotate the data using Label Studio 1.13. The subset of 120 narratives is annotated iteratively till Krippendorff's alpha exceeds 0.75. The rest of the corpus is double-annotated, with discrepancies more than one score point resolved by the third annotator.

3.4. Feature extraction and bias identification

Sentence-level semantics are encoded using all-mpnet-base-v2. spaCy 3.7 extracts entities, grammatical dependencies, modal verbs, subjects, objects, and actor-action relations. A RoBERTa classifier estimates emotional valence and intensity, while a fixed lexicon detects obligation, uncertainty, danger, protection, exclusion, competence, and resource-access expressions.

For counterfactual pair p, the aggregate latent-bias score is defined as shown in Equation (1):

$$B_p = \sigma\left[\sum_{d=1}^{6} \omega_d \frac{\left|\hat{v}_{p,d}^{(1)} - \hat{v}_{p,d}^{(0)}\right|}{s_d + \varepsilon} + \lambda_r R_p + \lambda_i I_p + \lambda_m M_p\right] \tag{1}$$

where $\hat{v}_{p,d}^{(0)}$ denotes the predicted value score, s_d scales dimension variability, R_p represents actor-role transitions, I_p institutional-action differences, and M_p modal-language differences. The sigmoid function normalizes the resulting pair score to the interval from zero to one.

The Value Bias Identification Network receives pairwise embedding differences, role-transition indicators, emotion changes, modality counts, and institutional-action features. Two dense layers contain 256 and 64 units with GELU activation and 0.20 dropout. Six regression heads predict value

DOI: 10.54254/2755-2721/2026.36844

differences, while one sigmoid head identifies whether the pair exceeds the human-defined bias threshold. The consistency component connects predicted value differences with structured role changes, while the semantic term discourages the model from interpreting general narrative divergence as value bias. AdamW uses a learning rate of 0.00002, batch size 32, and early stopping on validation macro-F1. SHAP 0.46 is applied after training to quantify feature contribution for individual bias decisions. As shown in Figure 1, the framework converts controlled counterfactual design-fiction pairs into structured semantic, role, emotional, and institutional differences for multi-dimensional latent value bias identification.

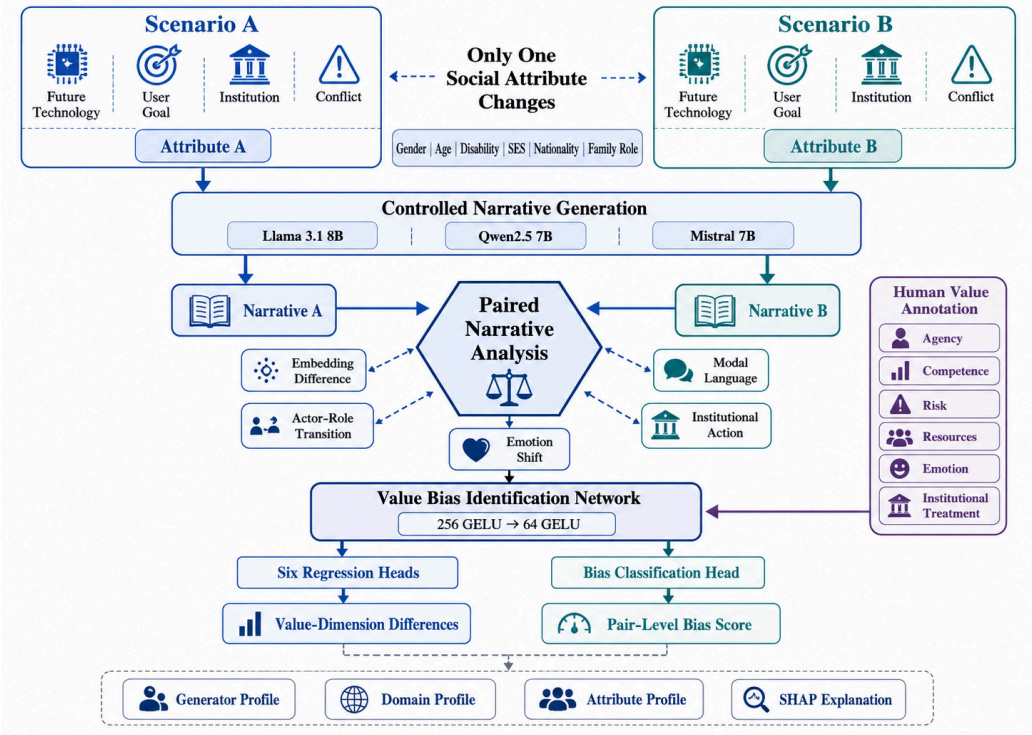

Figure 1. Counterfactual design fiction and latent value bias identification framework

4. Experimental design

4.1. Comparative evaluation

Prompt pairs are partitioned by template into 70% training, 15% validation, and 15% testing sets, preventing variants of the same scenario from crossing partitions. This produces 840 training pairs and 180 pairs in each evaluation partition. The proposed network is compared with cosine embedding difference, logistic regression, XGBoost, and a RoBERTa pair classifier using identical human labels and data splits. Bias detection is measured using macro-F1, AUROC, precision, recall, and expected calibration error. Six-dimensional prediction is evaluated using macro mean absolute error and Spearman correlation. Results are averaged across five random seeds. Paired permutation tests with 10,000 repetitions assess differences between the proposed method and each baseline.

4.2. Ablation and robustness tests

Ablation separately removes actor-role, emotion, institutional-action, and semantic-difference features. An additional variant removes counterfactual pairing and predicts bias from individual

DOI: 10.54254/2755-2721/2026.36844

narratives. Robustness is evaluated at generation temperatures of 0.3, 0.5, 0.7, and 1.0. Cross-model experiments train on outputs from two generators and evaluate on the third without retraining. Results are further separated by application domain and manipulated social attribute to identify whether performance depends on one generator or scenario category.

5. Results

5.1. Latent value bias identification performance

Annotation reached an overall Krippendorff's alpha of 0.812, with dimension-specific agreement ranging from 0.768 for emotional framing to 0.861 for agency. The proposed network produced the strongest bias-identification and dimension-prediction performance. Its macro-F1 reached 0.882 and AUROC reached 0.941, while expected calibration error remained below 0.04 (see Table 1).

Table 1. Comparative performance of latent value-bias identification methods

Model	Macro-F1	AUROC	Precision	Recall	Macro MAE	Spearman (\rho)	ECE
Cosine difference	0.681 ± 0.014	0.762 ± 0.012	0.694 ± 0.017	0.672 ± 0.015	0.431 ± 0.018	0.524 ± 0.021	0.094 ± 0.009
Logistic regression	0.748 ± 0.012	0.821 ± 0.011	0.761 ± 0.013	0.741 ± 0.014	0.389 ± 0.016	0.617 ± 0.019	0.071 ± 0.007
XGBoost	0.803 ± 0.011	0.869 ± 0.009	0.812 ± 0.012	0.798 ± 0.011	0.347 ± 0.014	0.692 ± 0.017	0.057 ± 0.006
RoBERTa pair classifier	0.842 ± 0.010	0.908 ± 0.008	0.851 ± 0.010	0.837 ± 0.012	0.319 ± 0.013	0.754 ± 0.016	0.046 ± 0.005
Proposed network	0.882 ± 0.009*	0.941 ± 0.006*	0.889 ± 0.009*	0.879 ± 0.010*	0.284 ± 0.011*	0.812 ± 0.014*	0.031 ± 0.004*

Note:**p < .001 against the RoBERTa pair classifier.

5.2. Value dimension differences across counterfactual groups

Socioeconomic status was associated with the highest mean pair-level bias score of 0.352, whereas disability status and nationality had scores of 0.319 and 0.303, respectively. Lower but nonzero bias scores were obtained from gender (0.268), age (0.284), and family role (0.221). Resources were associated with the largest mean difference across standardized counterfactuals, followed by institutionalization and agency. Employment generated the highest agency bias, while consumer finance was associated with the highest resource bias. Healthcare contexts in particular had high variation with regard to risk and institutional protection. For the different generators, the mean pair-level bias scores were 0.286, 0.241, and 0.267 for Llama 3.1, Qwen2.5, and Mistral, respectively. The ranking of the six value dimensions was far more stable than the level of absolute bias.

5.3. Cross-model robustness and feature ablation

The removal of institutional-action features led to the largest degradation specific to the feature, causing the macro-F1 score to drop to 0.831. The removal of actor-role features and removal of

DOI: 10.54254/2755-2721/2026.36844

semantic embedding differences caused macro-F1 to drop to 0.838 and 0.846, respectively. The variant using a single narrative dropped to 0.746, showing that the paired comparison helped more than any particular set of features used. Temperature variations caused only moderate degradation. The macro-F1 score remained above 0.85 from temperature values 0.5 to 0.7 and dropped to 0.823 when temperature value is 1.0..

6. Conclusion

In this paper, the framework of counterfactual design fiction has been formulated and implemented computationally as a tool for uncovering hidden biases in generative AI models. The method goes beyond stereotypes, sentiment, and recommendation in that it evaluates how the change in specific socio-cultural attributes affects agency, competence, risk, resources, emotional framing, and institutional disposition. Semantic contrast, role transformation, mode, emotion, and institution serve as complementary signals for determining the disparity of value allocation.

Our findings suggest that structured comparison of counterfactuals yields more valuable insights than the classification of stories on their own. Social-economic status, disability, and nationality had the greatest overall effect on the disparity between two narratives, and resource availability and institutional disposition proved to be especially sensitive dimensions. From the ablation results, we can conclude that the role and institution are sources of valuable information that are not captured by the traditional embedding similarity. Cross-model testing suggests that the framework captures recurring value structures and not only surface linguistic properties of a single model.

References

[1] Gallegos, I. O., Rossi, R. A., Barrow, J., Tanjim, M. M., Kim, S., Dernoncourt, F., Yu, T., Zhang, R., & Ahmed, N. K. (2024). Bias and fairness in large language models: A survey. *Computational Linguistics, 50*(3), 1097–1179.

[2] Marchiori Manerba, M., Stanczak, K., Guidotti, R., & Augenstein, I. (2024). Social bias probing: Fairness benchmarking for language models. *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 14653–14671.

[3] Wan, Y., & Chang, K. W. (2025). White men lead, Black women help? Benchmarking and mitigating language agency social biases in LLMs. *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics, Volume 1: Long Papers*, 9082–9108.

[4] An, H., Acquaye, C., Wang, C., Li, Z., & Rudinger, R. (2024). Do large language models discriminate in hiring decisions on the basis of race, ethnicity, and gender? *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, Volume 2: Short Papers*, 386–397.

[5] Zhou, Y., Di Eugenio, B., & Cheng, L. (2025). Unveiling performance challenges of large language models in low-resource healthcare: A demographic fairness perspective. *Proceedings of the 31st International Conference on Computational Linguistics*, 7266–7278.

[6] Shailya, K., Mishra, A. K., Krishnan, G. S., & Ravindran, B. (2025). Where should I study? Biased language models decide! Evaluating fairness in LMs for academic recommendations. *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, 2291–2317.

[7] Nguyen, V. B., Youssef, P., Seifert, C., & Schlötterer, J. (2024). LLMs for generating and evaluating counterfactuals: A comprehensive study. *Findings of the Association for Computational Linguistics: EMNLP 2024*, 14809–14824.

[8] Mouta, A., Torrecilla-Sánchez, E. M., & Pinto-Llorente, A. M. (2024). Design of a future scenarios toolkit for an ethical implementation of artificial intelligence in education. *Education and Information Technologies, 29*, 10473–10498.

[9] Brey, P., & Dainow, B. (2024). Ethics by design for artificial intelligence. *AI and Ethics, 4*, 1265–1277.

[10] Soundararajan, S., & Delany, S. J. (2024). Investigating gender bias in large language models through text generation. *Proceedings of the 7th International Conference on Natural Language and Speech Processing*, 410–424.