Fairness as a Creative Resource: Challenges and Opportunities in Creative Computing

Ricardo Trainotti Rabonato
Computer Science Department
Federal University of Sao Pauloy
Sao Jose dos Campos, SP, Brazil
trainotti.ricardo@unifesp.br

Lilian Berton
Computer Science Department
Federal University of Sao Paulo
Sao Jose dos Campos, SP, Brazil
lberton@unifesp.br

Abstract

Artificial Intelligence (AI) and generative models have expanded the landscape of computational creativity. While fairness is often framed as a corrective mechanism to mitigate social and cultural bias in AI systems, this paper explores an alternative perspective: fairness as a creative affordance capable of enhancing the aesthetic and cultural diversity of AI-generated content.

We propose a theoretical framework that repositions fairness as a potential catalyst for creativity and investigate this idea through a set of exploratory prompting experiments. Drawing on the few-shot learning capabilities of large language models, we examine whether minimal prompt modifications guided by fairness principles can lead to more diverse, inclusive, and contextually rich narratives.

Preliminary results suggest that fairness-oriented prompts may encourage generative systems to deviate from normative patterns, resulting in outputs with increased lexical variety, broader representation, and greater narrative complexity. While the study is limited in scale and scope, it opens new directions for designing generative systems that integrate ethical considerations as expressive constraints, fostering both innovation and cultural awareness in computational creativity.

Introduction

Creativity is a central theme in artificial intelligence, particularly in applications involving language, narrative, and symbolic expression. While large language models (LLMs) have made significant strides in generating fluent and coherent texts, guiding these systems toward outputs that are not only plausible but also original, expressive, and culturally situated remains a major challenge.

In the field of algorithmic fairness, efforts have traditionally focused on mitigating social and cultural biases in automated decision-making systems (Barocas, Hardt, and Narayanan 2023; Rabonato and Berton 2024). Fairness, in this view, operates primarily as a corrective mechanism. This paper proposes an alternative perspective: that fairness may also serve as a creative constraint — a guiding principle capable of expanding the narrative and aesthetic repertoire of generative models.

This proposition draws from theories in computational creativity that emphasize the generative role of constraints. Duch (2006) argues that creativity arises from the interaction between imagination and filtering mechanisms, whereby filters — including social, semantic, or aesthetic constraints — shape the emergence of novel combinations. From this viewpoint, fairness can act not only as a normative boundary, but as a stimulus for generating unexpected and meaningful associations.

The approach also aligns with Wiggins' (2006) formalization of Boden's concepts of exploratory and transformational creativity. Wiggins posits that a creative system can operate within a conceptual space (exploration) or challenge and expand the boundaries of that space (transformation). Fairness, as a constraint that destabilizes dominant narrative patterns, may thus operate as a trigger for transformational creativity — encouraging models to generate content that deviates from statistically dominant structures and engages with culturally diverse contexts (Wiggins 2006).

To explore this hypothesis, we conducted a set of small-scale prompting experiments involving narrative generation. Inspired by few-shot learning capabilities, we tested whether minimal modifications to input prompts — informed by fairness principles — could lead to more diverse and imaginative outputs. While the scope of this study is limited, preliminary results suggest that fairness-aware prompts may foster narrative complexity and symbolic richness, supporting the notion that ethical constraints can function as productive forces in creative AI systems.

Related Work

Fairness in Language Models

Recent research has investigated various strategies for mitigating bias in large language models (LLMs). Gallegos et al. (Gallegos et al. 2024) offer a comprehensive review and taxonomy of bias types and mitigation methods in NLP, while Doan et al. (Doan et al. 2024) synthesize recent advances with a focus on fairness considerations specific to the architecture and functioning of LLMs.

Despite progress in newer model versions, implicit biases remain persistent and highly sensitive to language, task formulation, and prompt structure (Torres et al. 2024). In this context, prompting techniques have emerged as a lightweight yet effective strategy for influencing model behavior. Zero-shot and few-shot prompting — including

chain-of-thought (CoT) reasoning — have been explored as means to induce fairness at inference time, without requiring model retraining (Zeng, Chung, and Zhou 2024). Other methods include architectural interventions and modifications to training objectives, such as fairness-aware loss functions (Wang, Li, and Zhang 2024; Atwood et al. 2024).

Bias and Computational Creativity

As LLMs are increasingly used in creative applications, questions about bias have gained relevance within the field of computational creativity. Traditionally, bias has been treated as an obstacle to be corrected, particularly from the standpoint of algorithmic fairness. However, some authors argue for a more nuanced perspective. Loughran (2022), for instance, suggests that biases — understood as tendencies, preferences, or learned associations — may be essential for creative cognition, acting as scaffolds for generating novelty amid otherwise conventional outputs (Loughran 2022).

This view intersects with broader discussions about the generative role of constraints in creativity. Constraints are often considered central to the emergence of innovation, especially when they provoke deviation from normative patterns (Lamb, Brown, and Clarke 2018). In this light, fairness can be interpreted not merely as a regulatory demand, but as a productive constraint that guides generative systems toward richer expressive spaces.

Brown and Ventura (2022) further argue that ethical and aesthetic dimensions can be co-encoded in computational creativity systems, with normative principles like fairness functioning as aesthetic filters that expand the range of acceptable outputs (Brown and Ventura 2022). Drawing on algorithmic information theory, they propose metrics such as logical depth and sophistication to assess creative value, suggesting that fairness constraints may contribute to increased complexity and originality in generative results.

Finally, Déguernel and Sturm (2023) conducted a systematic review of empirical studies on bias in the evaluation of computational creativity, concluding that human judgments are deeply shaped by sociocultural and contextual cues (Déguernel and Sturm 2023). Their findings support the idea that fairness-aware generation not only alters content characteristics, but also reshapes audience perception and reception — reinforcing the potential of fairness to function as both ethical and creative modulation.

Methodology

The experiments were conducted using a large language model (LLM), accessed through the free commercial version of the Claude model (Anthropic) via a web interface. It was not possible to adjust parameters such as temperature or output length. This limitation reinforces the focus on the semantic impact of the prompts and avoids technical interventions that could compromise the comparative analysis.

For this study, three sets of narrative prompt pairs were created. Each pair consists of a neutral version and a version with an explicit fairness orientation, by adding the instruction: "The narrative should show concern for justice and diversity." The prompt pairs used were as follows:

- **1a.** *Write a short story about an ordinary person who accidentally discovers a way to travel through time.*
- **1b.** *Write a short story about an ordinary person who accidentally discovers a way to travel through time. The narrative should show concern for justice and diversity.*
- **2a.** *Write a short story about a child who finds a mysterious object in their neighborhood.*
- **2b.** *Write a short story about a child who finds a mysterious object in their neighborhood. The narrative should show concern for justice and diversity.*
- **3a.** *Write a short story about a character attending a party in an unfamiliar place.*
- **3b.** *Write a short story about a character attending a party in an unfamiliar place. The narrative should show concern for justice and diversity.*

These prompts explore situations marked by the unexpected and the fantastic within everyday contexts, allowing for diverse narrative approaches. The first pair leans toward science fiction and adventure; the second evokes a playful tone, close to a childlike universe; and the third opens space for variations in style and atmosphere, ranging from the fantastic to the introspective, depending on the generative model's interpretation[1].

To assess creativity and ethical expressiveness, we followed the framework proposed by Lamb et al. (2018), which defines three central dimensions of creativity: *novelty*, *quality*, and *surprise*. These criteria served as the basis for the design of a structured evaluation instrument intended for future reader-based studies. The instrument comprises twelve items organized into four thematic blocks: (1) *fairness and representation*, (2) *creativity*, (3) *coherence and clarity*, and (4) *empathy and tone*.

The creativity block directly operationalizes the three core dimensions: originality of ideas and scenarios (*novelty*), presence of unexpected elements or plot twists (*surprise*), and formal-aesthetic consistency (*quality*). The additional blocks aim to capture ethical, affective, and structural dimensions of narrative impact, offering a broader perspective on cultural representation and reception.

Although no human evaluations were conducted at this stage, the instrument offers a foundation for future studies involving reader response. This choice reflects the interpretative nature of creativity, as discussed by Duch (2006) and Loughran (2022), and also acknowledges the limitations of automated metrics in capturing aesthetic and cultural nuance.

This approach aligns with the interpretative nature of creativity, as discussed by Duch (2006) and Loughran (2022), and also acknowledges the limitations of automated metrics in assessing aesthetic and cultural elements. The use of *fairness prompts* aims to provoke changes in the model's narrative behavior, encouraging outputs that are more expressive, elaborate, and less aligned with conventional patterns.

[1]The original texts generated from these prompts in Portuguese can be accessed at: `[https://bit.ly/supplementaryMaterial]`.

Discussion

The results suggest that simple interventions at the prompt level — specifically, the inclusion of fairness-oriented instructions — produce narratives that are not only longer, but also thematically richer, more diverse in character representation, and more attentive to social dynamics. This observation reinforces the central hypothesis of this work: that fairness, when framed not as a corrective post-processing constraint but as a generative input condition, can function as a creative catalyst. This aligns with previous arguments in the literature that emphasize the productive role of constraints in creative systems (Lamb, Brown, and Clarke 2018; Loughran 2022).

In the context of large language models (LLMs), this phenomenon can be interpreted as a semantic displacement within the model's latent space. By embedding ethical instructions into the prompt — such as a concern for justice and diversity — the generation process is nudged away from the high-frequency narrative templates statistically encoded in the training data. This shift disrupts the generative model's default behavior, reducing the influence of overrepresented cultural frames and opening space for less probable but potentially more imaginative combinations. In this sense, the prompt acts as a vector of conceptual divergence, prompting the model to access underutilized regions of its associative repertoire.

The observed increase in novelty and complexity is consistent with theories of computational creativity that view creative output as the product of navigating and transforming conceptual spaces (Wiggins 2006). By reframing fairness not as a filter applied to outputs, but as a transformation applied to the generative process itself, we reveal a new operational role for ethical constraints: as mechanisms of exploratory expansion. That is, fairness functions not only to prevent harm or bias, but to provoke narrative invention and aesthetic deviation from conventional paths.

Moreover, the interaction between fairness and creativity supports the notion of *constructive perturbation*. As the model integrates ethical concerns into narrative construction, it becomes more likely to introduce atypical characters, unexpected social contexts, and alternative resolutions — elements that contribute to the criteria of surprise and originality, two pillars of the creative framework adopted in this study. In technical terms, this suggests that fairness-oriented prompts modulate the decoding trajectory by activating latent associations that are otherwise suppressed under default optimization strategies.

This approach also resonates with the perspective of Brown and Ventura (2022), who argue that normative principles such as fairness can be encoded as aesthetic heuristics. In our results, prompts that asked for diversity and justice consistently resulted in outputs with greater emotional depth and cultural nuance. These findings suggest that ethical constraints can serve a dual function: guiding models toward socially responsible behavior and simultaneously enriching their expressive capacity.

From a methodological standpoint, this study contributes to ongoing debates on prompt engineering by demonstrating that minimal semantic interventions can yield disproportionately meaningful changes in output. This supports the viability of fairness-aware prompting as a lightweight, interpretable, and scalable strategy for shaping generative behavior — particularly in domains where creativity, cultural sensitivity, and user trust intersect.

It is important to note, however, that these findings are not generalizable across all genres, domains, or model families. The specific outcomes observed here depend on the interplay between the prompt, the model's training data, and the cultural priors embedded in its parameters. Future research should explore how different formulations of fairness (e.g., intersectionality, equity, inclusion) affect narrative generation, and whether these effects are consistent across languages, tasks, and audience expectations.

Finally, the results invite a broader reflection on the coevolution of ethical and creative capacities in generative AI. If fairness can indeed serve as a creative stimulus, then the development of ethically informed models need not be framed solely in terms of restriction or correction. Instead, it opens the possibility of reimagining AI creativity as an inclusive and dialogic practice — one that does not merely avoid harm, but actively contributes to the construction of richer, more plural symbolic worlds.

Exploratory Quantitative Analysis

To establish an initial comparative basis, we conducted a quantitative analysis of the generated texts, focusing on structural and surface-level features. Four metrics were considered: total word count, number of named characters, lexical variety (measured by the Measure of Textual Lexical Diversity, MTLD), and the number of identity or cultural diversity markers.

Named characters were identified using rule-based detection of proper nouns and character-referential phrases, manually verified to ensure accuracy. Diversity markers were annotated based on the presence of references to ethnicity, gender identity, religion, social class, age, and other cultural identifiers explicitly mentioned or strongly implied in the narrative.

MTLD, as described in (McCarthy and Jarvis 2010), is a metric designed to assess lexical diversity — the degree to which a text uses varied vocabulary. Unlike simpler measures such as the type-token ratio (TTR), MTLD is less sensitive to text length, making it more suitable for comparing texts of different sizes. It calculates the average number of words needed for the type/token ratio to reach a fixed threshold (typically 0.72). Higher MTLD scores indicate greater lexical variety.

Table 1 presents a summary of the analyzed texts.

While the scope of this pilot study is limited, some tentative patterns emerge. In each case, the fairness-oriented text (b) features more named characters, higher identity or cultural marker density, and — in most cases — greater lexical diversity. These patterns suggest a possible link between fairness-aware prompting and narrative expansion.

For example, in pair 1, *Seu Jonas's Clock* presents a higher MTLD (183.6 vs. 176.5), includes three named characters (versus one in *The Watchmaker*), and introduces five explicit diversity markers. While the first story centers on

Text	Words	Characters	MTLD	Markers
1a	597	1	176.5	0
1b	617	3	183.6	5
2a	364	2	135.5	1
2b	528	3	152.9	5
3a	464	1	146.7	1
3b	551	6	169.6	14

Table 1: Quantitative analysis and diversity markers in the generated stories

a generic protagonist in a neutral, introspective setting, the second is grounded in a social context, incorporating elements of inequality, memory, and community agency.

A similar contrast appears in pair 2: *The Time Medallion* develops a concise fantasy narrative with limited diversity (MTLD of 135.5; two characters; one marker), whereas *Sofia and the Pendulum of Time* presents a more elaborate plot with greater cultural engagement (MTLD of 152.9; three characters; five markers), addressing gentrification and historical repair.

Pair 3 exhibits the most pronounced difference: *The Party at the End of the World* is a metaphysical allegory with one character and a single abstract diversity reference, while *A Party in the Village of Colors* introduces six named characters and 14 diversity markers across multiple dimensions (ethnicity, religion, gender, age, etc.), achieving an MTLD of 169.6.

Although the numerical sample is small, these differences suggest that ethical prompting may encourage the model to activate a broader symbolic and social repertoire. Importantly, in fairness-aware texts, diverse characters frequently assume central narrative roles, such as protagonists, mentors, or agents of transformation — which may indicate not only increased representation but also richer plot structuring.

It is important to note that MTLD, while useful for indicating lexical variety, is not a direct measure of creativity. Future work will require more robust metrics — including semantic coherence, novelty, and human-rated creativity — to support stronger empirical claims. Nonetheless, these preliminary results reinforce the conceptual hypothesis that fairness can operate as a productive constraint, guiding the generation process toward more inclusive and narratively complex outputs.

Conclusion

This paper proposed a reframing of fairness in generative systems — not merely as a corrective mechanism against bias, but as a potential catalyst for creative diversity. Drawing on a set of exploratory narrative generation experiments, we observed that prompts incorporating principles of justice and inclusion tended to yield texts with more characters, greater lexical variety, and richer representations of cultural and social contexts.

Rather than constraining creativity, fairness-oriented prompts may function as productive perturbations, displacing the system from dominant narrative tropes and prompting engagement with less stereotypical symbolic and conceptual spaces. This aligns with perspectives in computational creativity that emphasize the generative potential of constraints, filters, and marginal frames.

Although the findings are preliminary and based on a small sample, they offer initial evidence that fairness can be operationalized as an expressive prompt design strategy. The results suggest a promising direction for developing generative systems that are not only more inclusive, but also more inventive in the narratives they construct.

Several limitations must be acknowledged. The study involved a limited number of prompts and model outputs, relied on surface-level and manually annotated metrics, and did not include human assessments of creativity or cultural resonance. Furthermore, while diversity markers were identified and discussed, deeper semantic and reception-based analyses remain necessary to avoid risks of superficial representation or stereotyping.

Future work should involve larger-scale evaluations, alternative creativity metrics, and engagement with diverse audiences to assess the impact of fairness-aware generation. Expanding the approach to other genres — such as poetry, scriptwriting, or speculative fiction — may also help uncover new affordances of ethical prompting as a driver of narrative innovation.

References

Atwood, J.; Scherrer, N.; Lahoti, P.; Balashankar, A.; Prost, F.; and Beirami, A. 2024. Inducing group fairness in prompt-based language model decisions.

Barocas, S.; Hardt, M.; and Narayanan, A. 2023. *Fairness and machine learning: Limitations and opportunities*. MIT press.

Brown, D., and Ventura, D. 2022. Ethics, aesthetics and computational creativity. In *ICCC*, 150–158.

Déguernel, K., and Sturm, B. L. T. 2023. Bias in Favour or Against Computational Creativity: A Survey and Reflection on the Importance of Socio-cultural Context in its Evaluation. In *International Conference on Computational Creativity*.

Doan, T. V.; Wang, Z.; Hoang, N. N. M.; and Zhang, W. 2024. Fairness in large language models in three hours. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, 5514–5517.

Gallegos, I. O.; Rossi, R. A.; Barrow, J.; Tanjim, M. M.; Kim, S.; Dernoncourt, F.; Yu, T.; Zhang, R.; and Ahmed, N. K. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics* 1–79.

Lamb, C.; Brown, D. G.; and Clarke, C. L. A. 2018. Evaluating computational creativity: An interdisciplinary tutorial. *ACM Comput. Surv.* 51(2).

Loughran, R. 2022. Bias and creativity. In *ICCC*, 354–358.

McCarthy, P. M., and Jarvis, S. 2010. Mtld, vocd-d, and hd-d: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods* 42(2):381–392.

Rabonato, R. T., and Berton, L. 2024. A systematic review of fairness in machine learning. *AI and Ethics* 1–12.

Torres, N.; Ulloa, C.; Araya, I.; Ayala, M.; and Jara, S. 2024. A comprehensive analysis of gender, racial, and prompt-induced biases in large language models. *International Journal of Data Science and Analytics* 1–38.

Wang, C.; Li, J.; and Zhang, R. 2024. A method to enhance structural fairness in large language models with active learning. *Authorea Preprints*.

Wiggins, G. A. 2006. Searching for computational creativity. *New Generation Computing* 24(3):209–222.

Zeng, C. C.; Chung, M.; and Zhou, E. 2024. Prompting for fairness: Mitigating gender bias in large language models with self-debiasing prompting. In *University of Michigan CSE 595 Natural Language Processing Fall 2024*.