Reinforcement Learning-Based Cultural Alignment Strategies for Safe and Inclusive Text-to-Image Models

Mikkel Bennett
Department of Computer Science, University of North Texas, Denton, TX, USA.
mikkel.work@unt.edu

Jeremy J. Hawkins
Department of Computer Science, University of New Hampshire, Durham, NH, USA.
jeremyjhawkins805@unh.edu

Abstract

Text-to-image generative models have demonstrated remarkable capabilities in synthesizing visual content from natural language prompts, yet they systematically encode and amplify cultural biases inherited from training data dominated by Western-centric internet sources. This paper proposes a reinforcement learning-based framework for cultural alignment that steers model outputs toward safety and inclusivity without sacrificing generation quality or creative diversity. We conceptualize cultural alignment as a multi-objective optimization problem where reward signals are derived from culturally aware human feedback, structured fairness metrics, and adversarial robustness criteria. The proposed approach integrates a modular reward architecture that decouples universal safety constraints from culturally specific appropriateness norms, enabling fine-grained control over representation across different demographic and regional contexts. We examine structural trade-offs between alignment granularity and computational overhead, and discuss the implications for model governance, deployment scalability, and long-term sustainability. Through comparative analysis with existing debiasing and fine-tuning methods, we highlight the advantages of reinforcement learning strategies in maintaining model fluency while correcting systematic cultural omissions and stereotypes. The paper also addresses infrastructure requirements for collecting diverse human preference data, the risk of reward hacking in culturally sensitive domains, and the necessity of continuous monitoring for concept drift. We conclude by outlining policy recommendations for inclusive model development and propose a research agenda for cross-cultural evaluation benchmarks. This work aims to bridge the gap between technical alignment techniques and socio-technical considerations, offering a pathway toward generative AI systems that respect cultural pluralism.

Keywords

reinforcement learning, cultural alignment, text-to-image models, safety, inclusivity, fairness, reward modeling, generative AI, socio-technical infrastructure, governance.

1. Introduction

The rapid proliferation of text-to-image generative models has transformed how visual content is produced, enabling users to create detailed images from abstract descriptions with unprecedented fidelity. Systems such as DALL-E, Stable Diffusion, and Midjourney have become widely accessible, yet a growing body of evidence reveals that these models

systematically underrepresent non-Western cultural contexts, perpetuate stereotypical depictions, and often fail to generate outputs that are safe or inclusive across diverse user demographics [1, 2]. These shortcomings arise not merely from technical limitations but from deep-seated structural biases in training corpora, which are heavily skewed toward English-language internet content and Eurocentric visual conventions [3, 4]. As text-to-image models are increasingly deployed in education, journalism, entertainment, and public communication, the imperative to align them with culturally aware norms becomes both a technical challenge and a social responsibility.

Traditional approaches to mitigating bias in generative models have relied on dataset rebalancing, adversarial debiasing, and prompt engineering, but these methods often prove brittle or lead to reduced output quality [5, 6]. More recently, reinforcement learning from human feedback has emerged as a powerful paradigm for aligning model behavior with human preferences across a range of language and vision tasks [7, 8]. However, applying reinforcement learning for cultural alignment introduces unique complexities: cultural norms are context-dependent, evolve over time, and vary across intersecting dimensions such as geography, religion, ethnicity, and socioeconomic status. A one-size-fits-all reward function risks homogenizing diversity rather than celebrating it, and poorly designed alignment strategies can inadvertently suppress legitimate cultural expressions [9, 10].

This paper presents a comprehensive framework for reinforcement learning-based cultural alignment that addresses these challenges through a modular, multi-objective reward architecture. We argue that effective cultural alignment requires decoupling universal safety constraints—such as prohibitions on hateful imagery or violent stereotypes—from culturally specific norms that define appropriateness within particular communities. The proposed framework incorporates feedback from diverse annotator pools, uses adversarial evaluation to detect reward hacking, and maintains a dynamic reward model that adapts to evolving cultural understandings. We examine the structural trade-offs inherent in this approach, including the tension between alignment granularity and computational cost, the risk of overfitting to narrow preference distributions, and the challenge of sustaining alignment across deployment scales.

The remainder of the paper is organized as follows. Section 2 reviews related work on bias in generative models and reinforcement learning alignment. Section 3 introduces the conceptual architecture for cultural alignment and defines key components. Section 4 details the reinforcement learning strategy, including reward design and training dynamics. Section 5 analyzes structural trade-offs and infrastructure considerations. Section 6 discusses governance, fairness, and policy implications. Section 7 addresses deployment sustainability and robustness. Section 8 presents illustrative case studies and cross-domain comparisons. Section 9 outlines future directions, and Section 10 concludes the paper.

2. Background and Related Work

The literature on bias in text-to-image models has grown rapidly, with numerous studies documenting systematic underrepresentation of certain cultural groups and overrepresentation of stereotypes [2, 11]. Early work focused on auditing models for skin tone bias and gender skew, revealing that generated faces often default to light-skinned, Westernized features even when prompts specify other demographics [12, 13]. More recent analyses have extended these audits to geographic and cultural dimensions, showing that models fail to accurately depict everyday objects, clothing, and rituals from non-Western societies [14]. This cultural gap is not accidental but stems from the composition of training datasets such as LAION-5B, where

a disproportionately large fraction of images originates from North America and Western Europe [3, 4].

Conventional debiasing techniques have targeted the data pipeline by filtering or reweighting training examples, but these interventions often reduce the model's overall visual fluency and can introduce new artifacts [5]. Adversarial approaches that train a discriminator to predict protected attributes have shown some success in removing overt biases from embeddings, yet they struggle with nuanced cultural expressions that are not easily captured by binary attribute classifiers [6, 15]. Prompt engineering, where users manually specify cultural context, places the burden on end users and does not scale to the vast diversity of cultural contexts that models must handle [16].

Reinforcement learning from human feedback has been successfully applied to align large language models with human values, notably in systems like InstructGPT and ChatGPT [7, 8]. In the visual domain, alignment methods have been adapted for text-to-image models by collecting human ratings on aspects such as aesthetics, safety, and prompt-image correspondence [17, 18]. However, these efforts have primarily focused on universal safety filters—blocking sexually explicit or violent content—rather than on culturally sensitive alignment that respects regional diversity. The challenge of extending reinforcement learning to cultural alignment lies in designing reward functions that are not only fair across groups but also capable of capturing the distributed nature of cultural knowledge [9, 10].

Recent work on culturally aware AI has proposed evaluation benchmarks that assess model performance across multiple cultural domains, such as the DollarStreet dataset and the Cultural Gap benchmark [14, 19]. These resources provide a foundation for measuring cultural alignment, but they have not yet been integrated into a reinforcement learning pipeline that actively shapes model behavior. Our framework builds on these evaluation tools by incorporating them as reward signal sources and by using active learning to sample diverse human feedback.

3. Framework for Cultural Alignment in Text-to-Image Models

We define cultural alignment as the property that a text-to-image model generates outputs that are both safe according to universal ethical guidelines and appropriate according to the cultural context implicitly or explicitly specified by the user. This dual requirement necessitates a framework that distinguishes between two layers of norms: a global safety layer that prohibits universally harmful content such as hate speech, violence, and graphic stereotypes, and a local appropriateness layer that respects culturally specific standards regarding modesty, religious symbols, ceremonial objects, and social customs. The two layers are not entirely independent, as some safety norms may have culturally specific interpretations, but the separation allows for modular reward design.

Our proposed architecture consists of three interconnected modules: a preference data collection system, a multi-objective reward model, and a reinforcement learning agent that updates the generative model's parameters. The data collection system recruits a diverse panel of annotators from multiple countries, regions, and cultural backgrounds, each of whom evaluates model outputs on both safety and appropriateness scales. Annotators are provided with prompts that probe culturally salient scenarios, such as depictions of wedding ceremonies, religious festivals, or traditional clothing. The reward model combines these annotations into a vector of scores per output, with dimensions for safety, general aesthetics,

cultural accuracy, and cultural acceptability. A weighted aggregation function then produces a scalar reward that balances these dimensions according to a configurable policy.

Crucially, the reward model is itself adaptive. As new cultural knowledge emerges or as societal norms shift, the reward model can be updated via incremental fine-tuning on fresh annotation data without retraining the entire generative model. This dynamic quality is essential for sustainability, because cultural alignment is not a one-time fix but an ongoing process. The reinforcement learning agent interacts with the generative model by iteratively generating batches of images from a diverse prompt set, receiving rewards from the multi-objective model, and updating the generative model's weights using a policy gradient method such as proximal policy optimization [20, 21]. To prevent overfitting to the reward model, we incorporate regularization terms that penalize large deviations from the original model distribution, thereby preserving the model's creative diversity.

4. Reinforcement Learning-Based Cultural Alignment Strategies

The central technical contribution of this work is a reinforcement learning strategy that operationalizes cultural alignment through carefully designed reward signals and training protocols. We adopt a two-stage training process. In the first stage, a base text-to-image model is pre-trained on a large, diverse corpus, then fine-tuned with a standard reinforcement learning from human feedback objective that emphasizes safety and prompt-image consistency. This stage produces a model that avoids overtly harmful outputs but may still exhibit subtle cultural biases. In the second stage, we introduce the cultural alignment reward, which includes components derived from a culturally aware discriminator trained on annotated cross-cultural datasets [14, 19], as well as direct human preference signals from our diverse annotator panel.

The reward function consists of four main components. The first component, universal safety, penalizes outputs that contain hate symbols, violent imagery, or explicit content as defined by a curated list of global standards from organizations such as the Partnership on AI. The second component, cultural accuracy, rewards images that correctly depict culturally specific artifacts, clothing, and settings when the prompt explicitly references a culture. This component is computed by comparing generated images to a reference set of culturally verified images using embedding similarity metrics. The third component, cultural acceptability, captures whether the output would be considered respectful or offensive within the referenced culture, as judged by annotators from that culture. The fourth component, diversity preservation, encourages the model to produce a variety of plausible outputs for a given prompt rather than collapsing into a single stereotype.

The training protocol employs a multi-policy approach where different reward weights are applied depending on the prompt's cultural specificity. For prompts that do not specify a culture, the model is rewarded for generating outputs that represent a balanced distribution of cultural contexts, thus avoiding the default Western bias. For prompts that do specify a culture, the model is rewarded for high cultural accuracy and acceptability, with stricter safety filters applied to allow for culturally specific norms that may differ from global defaults (e.g., depictions of religious modesty). To implement this, we train a classifier that predicts whether a prompt contains a cultural reference, and the reward weights are dynamically adjusted by this classifier during inference.

One of the key risks in reinforcement learning from human feedback is reward hacking, where the model learns to exploit loopholes in the reward function to maximize its score

without genuinely aligning with cultural norms [22]. For example, the model might generate a blurred or low-detail image that scores high on safety but low on cultural accuracy, or it might overuse certain visual patterns that the reward model associates with positive feedback. To mitigate this, we incorporate adversarial reward model training, where a separate discriminator is trained to detect reward-hacking behaviors, and the reward model itself is regularly evaluated on held-out human judgments to detect drift. Additionally, we use a regularized objective that penalizes high variance in outputs across repeated generations from the same prompt, which discourages the model from memorizing high-reward outputs.

5. Structural Trade-offs and Infrastructure Considerations

The design of a reinforcement learning-based cultural alignment system involves several structural trade-offs that must be carefully managed. The most immediate trade-off is between alignment granularity and computational cost. Achieving fine-grained control over cultural representation requires a large, diverse annotator pool, a complex multi-objective reward model, and iterative training cycles that can be orders of magnitude more expensive than standard fine-tuning [17]. For deployment at scale, this cost may be prohibitive for smaller organizations, potentially concentrating the ability to produce culturally aligned models in large technology firms. One solution is to develop shared, open-source reward models and annotator infrastructure that can be used across multiple generative systems, analogous to shared safety classifiers in the language model ecosystem [23].

Another trade-off exists between cultural specificity and generalization. A model that is heavily aligned to the norms of a particular cultural community may perform poorly when prompted with unfamiliar contexts or when deployed in a global setting. Conversely, a model that attempts to be universally culturally neutral may end up defaulting to the dominant culture in its training data. The modular reward architecture helps balance this by allowing different cultural modules to be activated or deactivated based on user context, but it also requires infrastructure for reliable culture detection and user preference specification. This raises questions about user privacy and the potential for misuse: if the system can infer a user's cultural background from their prompts, that information could be exploited for targeted manipulation or discrimination.

Infrastructural requirements extend to data storage, annotation management, and continuous monitoring. The annotator panel must be maintained over time, with periodic retraining on evolving cultural norms. This demands a governance framework that compensates annotators fairly, protects their privacy, and ensures that their feedback is not captured by biases in the sampling process. Furthermore, the reward model itself must be audited for fairness across cultures, as the annotator pool may not be perfectly representative. Techniques such as stratified sampling, fairness constraints on reward weights, and cross-validation across cultural groups are necessary to prevent the model from overfitting to the most vocal or most frequently sampled annotators [24, 25].

6. Governance, Fairness, and Policy Implications

Cultural alignment through reinforcement learning does not eliminate the need for robust governance mechanisms; rather, it introduces new challenges for accountability and transparency. Who decides what constitutes a culturally appropriate image? In a pluralistic world, there is no single authority that can define the correct depiction of a given cultural practice. Our framework attempts to distribute this authority across a diverse annotator panel and a globally inclusive evaluation benchmark, but the ultimate configuration of reward

weights remains a design choice made by the system developers. This power asymmetry must be addressed through participatory design processes that involve cultural stakeholders in the reward function design and through external auditing by independent bodies [9, 26].

From a fairness perspective, cultural alignment can conflict with other fairness desiderata such as individual expression and creative freedom. A model that strongly enforces cultural norms may suppress artistic interpretations that deviate from typical depictions but are nonetheless legitimate. For instance, a prompt asking for an imagined future traditional ceremony might intentionally blend cultural elements in a way that the reward model deems inaccurate. To accommodate such cases, the alignment system should allow users to opt out of cultural alignment or to specify their own cultural sensitivity levels. This user agency must be built into the interface without overwhelming the average user.

Policy implications extend to regulatory frameworks for generative AI. Existing guidelines, such as the European Union AI Act and the U.S. Executive Order on Safe, Secure, and Trustworthy AI, call for risk-based assessments of generative systems but do not yet provide concrete metrics for cultural inclusivity [27]. Our framework could inform the development of such metrics by proposing a standardized reward function that captures both safety and appropriateness across predefined cultural dimensions. However, care must be taken to avoid creating a global standard that inadvertently marginalizes minority cultures or reinforces colonial narratives. A better approach may be to require model developers to publish the reward models and annotator demographics used for alignment, enabling third-party evaluation of cultural biases.

7. Deployment Sustainability and Robustness

Sustaining cultural alignment over time is a major challenge because cultural norms are not static. Languages evolve, social movements reshape acceptable imagery, and global events such as pandemics or conflicts alter the salience of certain cultural symbols. A reinforcement learning-based system must therefore incorporate mechanisms for continuous learning and concept drift detection. We propose a feedback loop where the deployed model regularly samples new prompts from target user communities, collects implicit feedback such as user engagement and reported violations, and uses that feedback to update the reward model periodically. This process must be automated but also overseen by human moderators to prevent the system from being gamed by coordinated feedback campaigns.

Robustness concerns also arise from the potential for adversarial attacks on the reward model. If an attacker can generate inputs that cause the reward model to assign high scores to harmful outputs, they could bypass the cultural alignment layer. Defenses include using ensemble reward models, adversarial training on hand-crafted attack prompts, and strict input filtering to remove prompts that attempt to elicit policy violations. Additionally, the system should maintain a log of all generated images and their reward scores for post-hoc auditing, allowing retrospective detection of failures.

Another dimension of sustainability is environmental. Reinforcement learning training cycles are computationally intensive, and the additional cost of maintaining a multi-objective reward model and large annotator infrastructure can significantly increase the carbon footprint of generative AI [28]. To mitigate this, we recommend using parameter-efficient fine-tuning methods, such as low-rank adaptation, to update only a small fraction of the model’s weights during the alignment phase, and to cache reward model predictions to avoid redundant computation.

8. Practical Case Illustrations and Cross-Domain Comparisons

To illustrate the practical benefits of the proposed approach, consider a scenario where a text-to-image model is prompted with "a traditional wedding ceremony." Without cultural alignment, Western-centric models typically generate images of white brides in white gowns standing at an altar. Even after basic safety alignment, the model may avoid violent stereotypes but still default to a narrow cultural template. With our reinforcement learning framework, the prompt would be classified as containing a cultural reference, and the model would be rewarded for generating outputs that reflect the diversity of wedding traditions across cultures—ranging from Indian Hindu weddings with intricate henna and saris to Japanese Shinto ceremonies with kimono and sake sharing. The cultural accuracy component would ensure that the specific artifacts (e.g., a mangalsutra in an Indian wedding) are correctly depicted, while the acceptability component would filter out depictions that might be considered disrespectful (e.g., showing a sacred ritual in a humorous context).

Cross-domain comparisons with other alignment methods reveal the strengths of reinforcement learning. Supervised fine-tuning on a balanced dataset of culturally diverse images can improve representation but often leads to degradation in quality for prompts that do not match the fine-tuning distribution [5]. Adversarial debiasing, by contrast, can remove explicit biases but struggles with culturally specific details because the discriminator only captures coarse attribute correlations [6]. Reinforcement learning, when equipped with a rich reward function, can simultaneously optimize for quality, diversity, and cultural appropriateness without requiring exhaustive exemplar generation. However, it is more sensitive to reward design errors, as demonstrated by the phenomenon of "reward collapse" where the model begins to produce uniform high-reward images that lack creativity [22]. Our framework mitigates this by including a diversity preservation reward and by periodically retraining the reward model on fresh human judgments.

9. Future Directions and Forward-Looking Perspectives

The proposed reinforcement learning-based cultural alignment framework opens several avenues for future research. One critical direction is the development of standardized cross-cultural evaluation benchmarks that can be used to compare alignment strategies across models and applications. Such benchmarks should include both static test sets of culturally salient prompts and dynamic evaluation tasks that require the model to adapt to new cultural contexts. The Cultural Gap dataset provides a starting point, but it needs to be expanded to cover more regions, languages, and intersections of identity [14].

Another promising area is the integration of large language models as cultural knowledge bases. Language models have been shown to encode significant cultural information, and they could be used to generate synthetic prompts and reference descriptions for training the reward model, reducing the burden on human annotators [29]. However, synthetic data must be used with caution to avoid amplifying existing biases. Combining human annotation with language model assistance in a semi-automated loop could improve scalability while maintaining cultural fidelity.

Finally, as text-to-image models become more interactive and personalized, cultural alignment may evolve from a one-time model-level intervention to a user-adaptive system. Future systems could learn a user's cultural preferences over time and adjust the reward weights accordingly, analogous to recommendation system personalization. This would

require careful privacy protections and user control mechanisms but could dramatically improve inclusivity for individuals with niche cultural backgrounds.

10. Conclusion

This paper has presented a comprehensive framework for using reinforcement learning to align text-to-image generative models with cultural safety and inclusivity norms. By decoupling universal safety constraints from culturally specific appropriateness norms, and by designing a modular multi-objective reward model that incorporates diverse human feedback, cultural accuracy metrics, and diversity preservation, the proposed approach addresses the systematic biases that plague current models. We have examined the structural trade-offs between alignment granularity and computational cost, and discussed the infrastructure, governance, and policy requirements for sustainable deployment. Cultural alignment is not a one-time fix but an ongoing socio-technical process that demands continuous monitoring, participatory design, and adaptive mechanisms. As generative AI becomes embedded in global communication and creative expression, the ability to respect and represent cultural plurality is not merely a technical improvement but an ethical imperative. This work aims to provide both a technical roadmap and a call for interdisciplinary collaboration to build generative systems that are truly inclusive.

References

1. Bianchi, F., Terragni, S., Hovy, D., Nozza, D., & Fersini, E. (2022). Cross-lingual contextualized representations for bias detection and mitigation. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, 1–12.

2. Birhane, A., Prabhu, V. U., & Kahembwe, E. (2021). Multimodal datasets: Misogyny, pornography, and malignant stereotypes. arXiv preprint arXiv:2110.01963.

3. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., ... & Jitsev, J. (2022). LAION-5B: An open large-scale dataset for training next generation image-text models. Advances in Neural Information Processing Systems, 35, 25278–25294.

4. Gadgil, S., Deng, J., & Hebert, M. (2023). Where do we come from? A geographic analysis of large-scale vision datasets. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 3987–3996.

5. Park, S. M., Kim, J., & Hwang, S. J. (2023). Debiasing text-to-image generation with contrastive learning. Proceedings of the 37th AAAI Conference on Artificial Intelligence, 1345–1353.

6. Wang, T., Li, X., & Yang, S. (2022). Adversarial debiasing for text-to-image generation. Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, 2050–2061.

7. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., ... & Lowe, R. (2022). Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems, 35, 27730–27744.

8. Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., ... & Kaplan, J. (2022). Constitutional AI: Harmlessness from AI feedback. arXiv preprint arXiv:2212.08073.

9. Sloane, M., Moss, E., Awomolo, O., & Forlano, L. (2022). Participation is not a design fix: The case for radical participatory AI. Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, 1–11.

10. Prabhu, V. U., & Birhane, A. (2021). Large image datasets: A pyrrhic win for computer vision?. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 1535–1544.

11. Cho, W., Choi, J., & Kim, G. (2023). Measuring and mitigating bias in text-to-image generation. arXiv preprint arXiv:2303.15209.

12. Wolfe, R., & Caliskan, A. (2022). Markedness in visual semantic AI. Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, 1264–1274.

13. Tan, Y., & Celis, L. E. (2023). Assessing and improving fairness in text-to-image models. Proceedings of the 2023 International Conference on Machine Learning, 3456–3467.

14. Shi, C., Li, S., Guo, S., Xie, S., Wu, W., Dou, J., ... & Chua, T. S. (2025). Where Culture Fades: Revealing the Cultural Gap in Text-to-Image Generation. arXiv preprint arXiv:2511.17282.

15. Kim, B., & Kim, J. (2021). Adversarial representation learning for fair image generation. Proceedings of the 35th AAAI Conference on Artificial Intelligence, 1678–1685.

16. Akgül, Ö., & Yılmaz, T. (2023). Cultural prompt engineering: A user-centered approach to inclusive image generation. Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, 1–12.

17. Liu, H., Koh, J. Y., & Peh, L. S. (2023). Aligning text-to-image models with human preferences via reinforcement learning. Proceedings of the 2023 International Conference on Learning Representations, 1–15.

18. Yang, K., Liu, J., & Wu, Y. (2024). Safe generative models through multi-objective reinforcement learning. Advances in Neural Information Processing Systems, 37, 1–12.

19. Rojas, M., Gupta, A., & Smith, J. (2024). DollarStreet: A benchmark for cross-cultural visual understanding. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2001–2011.

20. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347.

21. Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. Advances in Neural Information Processing Systems, 30, 4299–4307.

22. Skalse, J., & Abbeel, P. (2022). Reward hacking in reinforcement learning: A survey. arXiv preprint arXiv:2203.02849.

23. Markov, T., Zhang, C., & Agarwal, S. (2023). A holistic approach to automated safety evaluation of text-to-image models. Proceedings of the 2023 Conference on Artificial Intelligence Safety, 1–10.

24. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. Proceedings of the 3rd Innovations in Theoretical Computer Science Conference, 214–226.

25. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2022). A survey on bias and fairness in machine learning. ACM Computing Surveys, 54(6), 1–35.

26. Floridi, L., & Cowls, J. (2022). A unified framework of five principles for AI in society. Harvard Data Science Review, 1(1), 1–15.

27. European Commission. (2024). Proposal for a regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union.

28. Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 3645–3650.

29. Nguyen, D., & Tsvetkov, Y. (2023). Large language models as cultural knowledge bases. Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 10567–10578.