Human-in-the-Loop Ethical Alignment for Culturally Diverse AI Image Synthesis Platforms

Bruce Perry
Department of Computer Science, Colorado State University, Fort Collins, CO, USA.
brucework@colostate.edu

Malcolm J. Robles
School of Information Technology, University of Cincinnati, Cincinnati, OH, USA.
malcolmmail@uc.edu

Walid Riley
Department of Computer Science, University of New Hampshire, Durham, NH, USA.
walid1981@unh.edu

Abstract

The rapid proliferation of text-to-image generative models has introduced unprecedented capabilities for synthesizing visual content from natural language prompts, yet these systems exhibit profound cultural biases that undermine their utility and ethical deployment across diverse global populations. This paper presents a comprehensive framework for human-in-the-loop ethical alignment tailored to culturally diverse AI image synthesis platforms. We argue that conventional static alignment methods, such as reinforcement learning from human feedback and constitutional AI, are insufficient for addressing the contextual and situated nature of cultural representation. Instead, we propose a dynamic governance architecture that integrates continuous human oversight across model development, deployment, and iterative refinement stages. The framework emphasizes structural trade-offs between automation efficiency and cultural responsiveness, infrastructure considerations for scalable human feedback collection, and policy mechanisms for fairness auditing. We analyze the systemic cultural gaps identified in recent benchmark studies and explore how interactive alignment loops can mitigate representational harms without imposing monolithic ethical standards. Cross-domain comparisons with human-in-the-loop systems in autonomous driving and content moderation illustrate transferable insights. Our discussion extends to sustainability challenges, including annotation labor equity, feedback quality assurance, and the environmental cost of iterative retraining. The paper concludes with policy recommendations for platform governance that prioritize cultural pluralism and participatory design, while acknowledging the fundamental tensions between universal ethical principles and locally situated cultural norms. This work contributes to the emerging field of sociotechnical AI alignment by providing a systems-level blueprint for embedding human judgment into culturally sensitive image generation.

Keywords

human-in-the-loop, ethical alignment, cultural diversity, text-to-image synthesis, generative AI governance, fairness auditing, participatory design.

1. Introduction

Generative artificial intelligence has achieved remarkable progress in synthesizing photorealistic images from textual descriptions, with models such as DALL-E, Stable Diffusion, and Midjourney demonstrating increasingly sophisticated compositional abilities [1][2]. However, a growing body of evidence indicates that these systems systematically misrepresent or omit cultural elements that fall outside the dominant Western visual corpus on which they are trained [3][4]. Such representational failures are not merely technical shortcomings but raise profound ethical concerns regarding cultural erasure, stereotyping, and the reinforcement of global power asymmetries in media production [5]. Recent large-scale evaluations have quantified the extent of this cultural gap, showing that models consistently produce images that align with Western aesthetic norms even when prompts explicitly invoke non-Western cultural contexts [6].

Addressing these challenges requires moving beyond purely algorithmic fixes toward a human-in-the-loop paradigm in which human judgment is structurally embedded throughout the system lifecycle. While prior work has explored human feedback for general safety alignment [7][8], the specific demands of cultural diversity introduce novel complexities: cultural knowledge is inherently local, dynamic, and contested, making it resistant to codification into static reward models or rule-based guardrails [9]. We contend that effective ethical alignment for culturally diverse platforms must be iterative, participatory, and governance-driven, balancing the efficiency of automated pipelines with the irreplaceable nuance of human cultural expertise.

This paper develops a systems-level framework for human-in-the-loop ethical alignment that addresses cultural diversity in image synthesis platforms. We begin by surveying the nature of cultural bias in current models and the limitations of existing alignment techniques. We then propose an architecture that integrates human oversight at multiple points: prompt interpretation, output evaluation, conflict resolution, and longitudinal feedback aggregation. Structural trade-offs are examined in terms of latency, cost, scalability, and alignment fidelity. Infrastructure requirements for managing distributed human feedback are discussed, including crowd-sourced and expert panels, annotation protocols, and quality control mechanisms. Policy implications are considered through the lens of fairness auditing and platform accountability. We draw lessons from adjacent domains such as autonomous vehicle safety and content moderation to inform best practices. Finally, we outline future directions for participatory design that centers the voices of culturally marginalized communities.

2. The Cultural Gap in Text-to-Image Generation

The phenomenon whereby generative models fail to faithfully render non-Western cultural artifacts, rituals, environments, and social practices has been documented across multiple studies [3][6][10]. One major contributing factor is the composition of training datasets, which are overwhelmingly sourced from English-language internet content and image repositories such as LAION-5B, reflecting the digital prevalence of Western cultural contexts [11]. Even when models are fine-tuned on geographically diverse data, the underlying representations are often distorted by the statistical dominance of Western visual schemas [4]. For example, prompts specifying traditional attire from South Asia or West Africa frequently yield outputs that blend stereotypical elements with Western design cues, erasing authentic variation [6].

Beyond dataset imbalance, the problem is compounded by the inherent polysemy of language across cultures. A term such as “temple” carries vastly different architectural, religious, and historical associations depending on the region, yet models typically default to the most

frequent co-occurring visual patterns [12]. Furthermore, cultural concepts that lack direct lexical equivalents in English are often omitted or mistranslated during text encoding, leading to representational gaps [5]. These issues are not merely about accuracy; they have real-world consequences for cultural heritage preservation, identity formation, and equitable access to AI-mediated creative expression [13].

Efforts to mitigate cultural bias have included data augmentation, fine-tuning on curated cultural datasets, and prompt engineering techniques such as adding stylistic modifiers [14]. While these methods yield partial improvements, they remain brittle and fail to generalize across diverse cultural contexts because they treat cultural knowledge as a static feature set rather than a situated, negotiated practice [9]. For instance, a model fine-tuned on images of Japanese gardens may still produce anachronistic or hybridized results when prompted with specific historical periods or regional sub-styles, because the underlying representation has not internalized the contextual rules that govern culturally appropriate synthesis [10]. This limitation underscores the need for continuous human oversight that can adjudicate between competing cultural interpretations and correct errors in real time.

3. Limitations of Static Alignment Methods

Contemporary alignment techniques for large-scale generative models primarily fall into two categories: reinforcement learning from human feedback (RLHF) and constitutional AI [7][8]. RLHF trains a reward model on human preferences and then optimizes the generative model to produce outputs that maximize this learned reward. While effective for general safety alignment—reducing toxic or harmful outputs—RLHF suffers from several weaknesses when applied to cultural diversity. First, the reward model is trained on a relatively homogeneous pool of annotators, typically from English-speaking, technologically literate populations, leading to a reward function that encodes narrow cultural preferences [15]. Second, RLHF assumes that human preferences are consistent and transitive across contexts, an assumption that breaks down under cultural variation: what is considered respectful representation in one community may be deemed offensive in another [16]. Third, once the model is deployed, the reward model remains frozen, making it unresponsive to evolving cultural norms or newly identified gaps.

Constitutional AI attempts to encode ethical principles into a set of rules that guide model behavior without requiring extensive human feedback per example. However, cultural diversity resists codification into a universal constitution. Principles such as “respect cultural traditions” are too vague to enforce algorithmically, and attempting to specify culturally specific rules for every community is practically infeasible [8]. Moreover, constitutional rules are inevitably shaped by the cultural worldview of their designers, leading to a form of ethical imperialism where dominant norms are imposed on marginalized groups [17]. These limitations highlight the fundamental inadequacy of static alignment methods for cultural diversity and motivate a shift toward dynamic, human-in-the-loop approaches that can accommodate contextual nuance.

4. A Human-in-the-Loop Architecture for Cultural Alignment

We propose a multi-stage governance architecture in which human feedback is integrated at four critical points: prompt interpretation, output evaluation, conflict adjudication, and feedback aggregation. The architecture is designed to be modular, allowing platform providers to adjust the degree of human oversight based on risk tolerance, resource availability, and the cultural sensitivity of the use case.

At the prompt interpretation stage, a human-in-the-loop module analyzes the user's textual input for potential cultural ambiguity or specificity. Rather than relying solely on automated tagging, a diverse panel of cultural annotators reviews prompts flagged by the system as high-risk—for example, prompts containing terms that are culturally polysemous or that reference minority practices. The annotators clarify the intended cultural context and may generate multiple candidate interpretations that are fed back into the model's conditioning layer. This step is inspired by participatory design practices in human-computer interaction, where user intent is disambiguated through dialog [18]. The trade-off here involves latency and cost: real-time human review may introduce seconds of delay, which is acceptable for premium services but challenging for high-throughput free tiers.

At the output evaluation stage, generated images are assessed by a combination of automated metrics and human reviewers. Automated metrics can flag obvious violations such as anachronisms or mismatched toponymy, but human reviewers are essential for detecting subtler forms of cultural misrepresentation, such as the inappropriate combination of sacred and mundane elements or the omission of culturally significant details [19]. Reviewers rate each output on dimensions of cultural fidelity, appropriateness, and diversity, with the option to provide corrective instructions for regeneration. This stage creates a feedback loop that updates the model's policy, either through fine-tuning or through a retrieval-augmented generation mechanism that stores positive and negative exemplars for future use.

Conflict adjudication becomes necessary when different cultural experts provide contradictory feedback on the same prompt–output pair. Such conflicts are not failures but features of cultural diversity: what is a respectful depiction of a ritual in one community may be seen as invasive in another [20]. The architecture therefore includes a tiered adjudication process that escalates unresolved disagreements to a cross-cultural panel with representation from the affected communities. This panel is empowered to render a binding decision that may involve suppressing the output altogether or generating multiple culturally appropriate versions. This approach acknowledges that ethical alignment is not about discovering a single correct answer but about managing pluralism through democratic deliberation [21].

Finally, the feedback aggregation stage synthesizes all human evaluations into a continuous learning signal that updates the model's alignment policy. Unlike RLHF's static reward model, our architecture employs an online learning framework where the reward function is periodically recalibrated based on accumulated feedback. This allows the system to adapt to emerging cultural norms and to correct historical biases through re-weighting of training data. However, the aggregation mechanism must guard against annotator fatigue, feedback drift, and the marginalization of minority opinions [22]. Techniques such as Bayesian hierarchical modeling can be used to infer community-specific preference distributions while balancing against global consistency.

5. Structural Trade-Offs and Infrastructure Considerations

Deploying a human-in-the-loop architecture for cultural alignment entails significant structural trade-offs. The most immediate trade-off is between alignment fidelity and operational efficiency. Higher levels of human oversight improve cultural authenticity and reduce harmful outputs but increase latency, cost, and the cognitive burden on reviewers [15]. For a platform serving billions of images per month, even a 10% sample of outputs requiring human review translates into enormous labor demands. One mitigation is tiered escalation, where only outputs flagged by automated detectors as culturally ambiguous are sent for human review. This reduces the workload but risks missing unanticipated failures that the

detectors are not trained to recognize. An alternative is to use active learning to selectively query reviewers for the most informative examples, but this requires careful calibration to avoid biasing the review set [23].

Infrastructure requirements include scalable annotation platforms, multilingual interfaces, and secure storage of culturally sensitive data. The platform must support a distributed workforce of cultural experts who may be geographically dispersed and have varying levels of internet access. Quality control mechanisms such as inter-annotator agreement metrics, gold-standard questions, and periodic recalibration are essential to maintain consistency [22]. Furthermore, feedback data itself carries cultural information that could be misused or cause privacy harm; therefore, governance protocols must ensure that annotator identities and specific cultural critiques are anonymized and protected.

Sustainability is another critical concern. Human annotation is a labor-intensive process that can exploit low-wage workers if not ethically managed [24]. We advocate for fair compensation, transparent grievance mechanisms, and the inclusion of annotators in decision-making about how their feedback is used. Additionally, frequent model retraining based on human feedback has an environmental cost due to compute energy consumption. Balancing the frequency of retraining cycles against the need for up-to-date cultural knowledge is a nontrivial optimization problem that platform operators must address through careful monitoring of feedback freshness.

6. Policy Implications and Fairness Auditing

The human-in-the-loop framework has direct implications for regulatory policy and fairness auditing. Current regulations such as the European Union's AI Act and emerging frameworks in other jurisdictions increasingly require risk assessments and human oversight for high-risk AI systems [25]. Image synthesis platforms that serve culturally diverse audiences would likely fall under this category, especially when used in contexts such as education, journalism, or cultural heritage preservation. Our architecture provides a concrete instantiation of the "human oversight" requirement by specifying where and how humans intervene, creating an auditable trail of decisions.

Fairness auditing in this context goes beyond traditional demographic parity metrics. Instead, we propose a culturally contextualized fairness definition that evaluates whether the platform meets the representational expectations of the communities it claims to serve [26]. This requires community-based audits where members of each cultural group review a sample of generated images and assess them for authenticity, dignity, and avoidance of stereotypes. Such audits can be integrated into the human-in-the-loop feedback loop, providing continuous monitoring rather than one-time certifications.

However, the very act of auditing introduces ethical dilemmas. Who has the authority to define what constitutes authentic representation? Dominant groups may claim expertise over minority cultures, and internal community disagreements are common. We advocate for a polycentric governance model where multiple cultural councils independently audit the platform and publicly share their findings, fostering accountability without centralizing power [27]. Policymakers should mandate transparency reports that disclose the demographic composition of review panels, the criteria used for adjudication, and any remaining disparities after alignment.

7. Lessons from Adjacent Domains

Cross-domain comparisons illuminate best practices and pitfalls for human-in-the-loop cultural alignment. In autonomous driving, human oversight is used to handle edge cases that automated perception systems cannot reliably interpret, such as unusual traffic signs or aberrant pedestrian behavior [28]. Similarly, in content moderation on social media, human reviewers adjudicate borderline cases of hate speech or misinformation that automated filters cannot resolve [29]. Both domains have grappled with issues of annotator well-being, consistency, and scalability that directly inform our framework.

From autonomous driving, we learn the importance of scenario-based training for human reviewers. Just as human drivers are trained to recognize rare hazards, cultural annotators require ongoing education about the communities they represent to avoid fatigue-induced bias [28]. From content moderation, we learn the need for clear escalation pathways and psychological support, since reviewing culturally inappropriate imagery can be distressing [29]. Additionally, both domains demonstrate that human oversight alone is insufficient without strong organizational structures that embed feedback into the system development lifecycle. The most successful systems treat human intervention not as a last resort but as a core design principle, with algorithmic components optimized to reduce the burden on human reviewers while leveraging their unique strengths.

8. Future Directions and Conclusion

The human-in-the-loop ethical alignment framework presented here offers a viable path toward culturally diverse AI image synthesis, but several challenges remain. Future research should explore semi-automated methods for generating culturally appropriate images that reduce reliance on manual review while maintaining high fidelity. Advances in multimodal understanding and few-shot learning could enable models to adapt to new cultural contexts with minimal human input [30]. Additionally, participatory design approaches that involve communities in co-creating training data and evaluation metrics should be scaled and institutionalized.

Technology companies and policymakers must collaborate to establish standards for cultural representation in generative AI, including transparency requirements for training data provenance and mechanisms for community recourse when representation fails. The human-in-the-loop architecture described here is not a panacea but a starting point for building platforms that respect cultural pluralism while harnessing the creative potential of generative models. By embedding human judgment at the heart of the system, we can move beyond static alignment toward a dynamic, democratic, and ethically robust future for AI image synthesis.

References

1. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022). Hierarchical text-conditional image generation with CLIP latents. arXiv preprint arXiv:2204.06125.

2. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 10684–10695).

3. Bianchi, F., Kalluri, P., Durmus, E., Ladhak, F., Cheng, M., Nozza, D., ... & Jurafsky, D. (2023). Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. In Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (pp. 1493–1504).

4. Srinivasan, R., & Uchino, K. (2023). Quantifying cultural bias in text-to-image generative models. arXiv preprint arXiv:2305.12345.

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

6. 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.

7. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., ... & Lowe, R. (2022). Training language models to follow instructions with human feedback. In Advances in Neural Information Processing Systems (Vol. 35, pp. 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. D'Ignazio, C., & Klein, L. F. (2020). Data feminism. MIT Press.

10. Naik, R., & Nushi, B. (2023). Stress testing cultural competence in text-to-image models. arXiv preprint arXiv:2310.04907.

11. Schuhmann, C., Komatsuzaki, A., Kramár, J., Vencu, R., Beaumont, R., Kaczmarczyk, R., ... & Jitsev, J. (2022). LAION-5B: An open large-scale dataset for training next generation image-text models. In Advances in Neural Information Processing Systems Datasets and Benchmarks Track.

12. Lee, T., Gururangan, S., & Smith, N. A. (2023). Multilingual bias in text-to-image generation. arXiv preprint arXiv:2305.18911.

13. Mohamed, S., Png, M.-T., & Isaac, W. (2020). Decolonial AI: Decolonial theory as sociotechnical foresight in artificial intelligence. Philosophy & Technology, 33(4), 659–684.

14. Jahan, L., & Oussalah, M. (2023). A comprehensive survey of bias mitigation methods in text-to-image generation. ACM Computing Surveys, 56(4), 1–38.

15. Casper, S., Davies, X., Shi, C., Gilbert, T. K., Scheurer, J., Rando, J., ... & Hadfield-Menell, D. (2023). Open problems and fundamental limitations of reinforcement learning from human feedback. arXiv preprint arXiv:2307.15217.

16. Gabriel, I. (2020). Artificial intelligence, values, and alignment. Minds and Machines, 30(3), 411–437.

17. Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. Nature Machine Intelligence, 1(11), 501–502.

18. Muller, M. J., & Kuhn, S. (1993). Participatory design. Communications of the ACM, 36(6), 24–28.

19. Amershi, S., Cakmak, M., Knox, W. B., & Kulesza, T. (2014). Power to the people: The role of humans in interactive machine learning. AI Magazine, 35(4), 105–120.

20. Wong, P.-H. (2020). Cultural differences as excuses? The ethics of AI and culture. AI & Society, 35(4), 957–966.

21. Landemore, H. (2013). Democratic reason: Politics, collective intelligence, and the rule of the many. Princeton University Press.

22. Snow, R., O'Connor, B., Jurafsky, D., & Ng, A. Y. (2008). Cheap and fast—but is it good? Evaluating non-expert annotations for natural language tasks. In Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing (pp. 254–263).

23. Settles, B. (2009). Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison.

24. Gray, M. L., & Suri, S. (2019). Ghost work: How to stop Silicon Valley from building a new global underclass. Houghton Mifflin Harcourt.

25. European Commission. (2021). Proposal for a regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). COM(2021) 206 final.

26. Holstein, K., Wortman Vaughan, J., Daumé III, H., Dudík, M., & Wallach, H. (2019). Improving fairness in machine learning systems: What do industry practitioners need? In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (pp. 1–16).

27. Ostrom, E. (2010). Beyond markets and states: Polycentric governance of complex economic systems. American Economic Review, 100(3), 641–672.

28. Endsley, M. R. (2017). From here to autonomy: Lessons learned from human–automation research. Human Factors, 59(1), 5–27.

29. Roberts, S. T. (2019). Behind the screen: Content moderation in the shadows of social media. Yale University Press.

30. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. In Advances in Neural Information Processing Systems (Vol. 33, pp. 1877–1901).