Cross-Cultural Explainability Metrics for Evaluating Ethical Compliance in AI-Generated Visual Content

Rohan C. Parekh
Department of Computer Science, University of Central Florida, Orlando, FL, USA.
hellorohan@ucf.edu

Christopher Lewis
Department of Computer Science and Engineering, University of Nevada, Reno, Reno, NV, USA.
christopher.lewis270@unr.edu

Yun Liu
Department of Computer Science, University of Alabama at Birmingham, Birmingham, AL, USA.
yunmail@uab.edu

Abstract

The rapid proliferation of generative artificial intelligence systems capable of producing photorealistic visual content has introduced profound challenges for ethical compliance across culturally heterogeneous user populations. Existing explainability frameworks, largely developed within Western epistemological traditions, often fail to account for the diverse normative expectations, interpretative schemas, and value systems that shape how individuals perceive and evaluate AI-generated imagery. This paper proposes a systematic framework for cross-cultural explainability metrics tailored to the evaluation of ethical compliance in visual content generation. We argue that ethical compliance cannot be reduced to universal checklists but must be operationalized through metrics that are sensitive to culturally variable constructs such as fairness, dignity, authenticity, and harm. Our framework integrates insights from comparative philosophy, socio-technical systems theory, and explainable AI research to define a multi-dimensional metric space comprising representational accuracy, contextual transparency, normative alignment, and user-centered intelligibility. We examine structural trade-offs between global standardization and local adaptation, architectural considerations for embedding culturally aware explainability components into generative pipelines, and governance implications for platform accountability and content moderation. Through cross-domain comparisons with healthcare AI and autonomous systems, we illustrate the generalizability and limitations of the proposed approach. The paper further addresses robustness and sustainability challenges, including the mitigation of feedback loops that reinforce cultural stereotypes and the long-term maintenance of culturally responsive explanation mechanisms. Policy implications are discussed with recommendations for regulatory frameworks that mandate cross-cultural explainability audits. We conclude by outlining a research agenda for developing dynamic, community-informed metrics that evolve with shifting cultural landscapes.

Keywords

cross-cultural explainability, ethical compliance, AI-generated visual content, fairness metrics, socio-technical systems, cultural alignment.

1. Introduction

The deployment of large-scale generative models for visual content creation, such as text-to-image and image-to-image systems, has accelerated dramatically in recent years, enabling unprecedented capabilities in design, entertainment, advertising, and education. However, these systems frequently produce outputs that reflect the biases and cultural blind spots embedded in their training data, leading to representations that may be ethically problematic when viewed through diverse cultural lenses. A single image generated by a model can be interpreted as innocuous in one cultural context and deeply offensive in another, raising urgent questions about how to evaluate and enforce ethical compliance at a global scale. Explainability, the ability of a system to provide intelligible justifications for its outputs, has emerged as a cornerstone of trustworthy AI, yet existing explainability methods are overwhelmingly designed for Western, English-speaking, and individualistic cultural settings. This paper contends that without culturally adaptive explainability metrics, efforts to ensure ethical compliance in AI-generated visual content will remain incomplete and potentially harmful.

The motivation for cross-cultural explainability metrics stems from recognition that ethical norms are not universal but are negotiated within specific historical, religious, and social frameworks. For example, notions of modesty, respect for authority, and collective honor vary widely across East Asian, South Asian, Middle Eastern, African, and Indigenous cultures, and these variations directly affect how visual content should be evaluated. An explainability metric that measures only factual accuracy or feature attribution fails to capture whether an AI system has respected culturally specific taboos, such as the depiction of sacred symbols, ancestral figures, or gender roles in ways that align with community standards. Thus, a narrow technical definition of explainability must be expanded to encompass cultural intelligibility, that is, the degree to which an explanation is not only accurate but also meaningful and contextually appropriate for a given audience.

This paper advances a systems-level perspective that integrates explainability metrics into the broader infrastructure of generative AI deployment. We consider the entire pipeline, from training data curation and model architecture to output generation, user interaction, and feedback loops. The proposed framework does not prescribe a single set of metrics but rather a meta-framework for designing culturally adaptive explainability scores that can be calibrated to different regions and communities. We examine structural trade-offs between computational efficiency, scalability, and cultural fidelity, and we explore how governance mechanisms such as auditing protocols, participatory design, and regulatory sandboxes can support the adoption of such metrics. The remainder of the paper is organized as follows: Section 2 reviews related work on cultural bias in AI and explainability. Section 3 develops a theoretical framework for cross-cultural explainability. Section 4 proposes specific metric classes and discusses their structural trade-offs. Section 5 addresses architectural and governance implications. Section 6 provides cross-domain illustrations. Section 7 examines robustness, fairness, and sustainability. Section 8 discusses policy and deployment pathways. Section 9 concludes with a research agenda.

2. Background and Related Work

The problem of cultural bias in AI systems has received growing attention, particularly in natural language processing and computer vision. Bender et al. [1] demonstrated how large language models can amplify hegemonic worldviews when trained on predominantly English texts. Similarly, in visual generation, Liang et al. [2] showed that text-to-image models

associate certain professions with specific ethnicities in ways that reinforce stereotypes. The required reference by Shi et al. [6] revealed a systematic cultural gap in text-to-image generation, wherein visual outputs default to Western-centric interpretations of culturally situated prompts. These findings underscore the need for evaluation metrics that explicitly account for cultural context rather than assuming a universal baseline.

Explainable AI (XAI) has developed a rich portfolio of methods, including saliency maps, concept activation vectors, and counterfactual explanations. Doshi-Velez and Kim [3] provided a foundational taxonomy of interpretability evaluation, distinguishing between application-grounded, human-grounded, and functionally grounded approaches. Miller [4] further argued that explanations are social phenomena and that AI explanations must conform to cognitive and conversational norms. However, these works largely assume a homogeneous user base with shared expectations about what constitutes a good explanation. In contrast, cross-cultural psychology, as reviewed by Hofstede [5] and Markus and Kitayama [7], demonstrates that individuals from different cultures prefer different explanatory styles: for instance, East Asians tend to favor holistic, situational explanations, while Westerners prefer analytic, dispositional ones. Translating these insights into measurable explainability metrics for visual content remains an open challenge.

Ethical compliance frameworks for AI, such as the OECD AI Principles [8] and the EU AI Act risk categories [9], emphasize transparency, accountability, and fairness. However, these frameworks rarely specify how to operationalize fairness across cultures. Mehrabi et al. [10] provided a comprehensive survey of fairness definitions and bias mitigation techniques, yet cultural fairness is often treated as a subcategory of demographic fairness, ignoring the fact that cultural groups are not merely demographic categories but embedded in historically evolved meaning systems. Gebru et al. [11] called for standardized documentation practices (e.g., datasheets) to improve accountability, but such documentation is typically designed from the developer's perspective rather than from the end user's cultural standpoint. Our work builds on these foundations by proposing metrics that bridge the gap between technical explainability and culturally situated ethical evaluation.

3. Theoretical Framework for Cross-Cultural Explainability

A cross-cultural explainability metric must be grounded in a theoretical understanding of how explanation functions within a socio-technical system. We adopt a constructivist perspective in which meaning is co-created through interaction between the AI system, the generated visual content, and the culturally embedded user. Explanation is not merely a transmission of information but a communicative act that must achieve relevance, coherence, and legitimacy according to the recipient's interpretive community. Our framework comprises four dimensions: representational accuracy, contextual transparency, normative alignment, and user-centered intelligibility.

Representational accuracy measures the degree to which the generated visual content faithfully corresponds to the intended semantics without introducing culturally incongruent elements. For example, if a user prompts for a "traditional wedding ceremony," the system should not default to white wedding gowns if the intended culture typically uses red garments. This dimension relies on culturally annotated datasets and participatory validation. Contextual transparency assesses the extent to which the system provides information about the sources, assumptions, and limitations that influenced the output. A culturally transparent explanation might include metadata about the training data demographics, the cultural provenance of visual features, or the degree of uncertainty regarding cultural alignment. Normative

alignment evaluates whether the output respects the explicit and implicit ethical norms of the target culture. This dimension is inherently contestable because norms are dynamic and vary within cultures. Thus, normative alignment metrics must be continuously updated through community feedback mechanisms. User-centered intelligibility measures how comprehensible and actionable the explanation is to the end user. For a user who is not an AI expert, an explanation should avoid technical jargon and instead use culturally familiar analogies or narratives.

These four dimensions interact in complex ways. A high level of representational accuracy may still yield low normative alignment if the cultural norm itself is contested. Similarly, contextual transparency can be overwhelming if it provides too much information without considering the user's cognitive load. The trade-off between comprehensiveness and intelligibility is a central design challenge. Our framework does not prescribe weights for these dimensions but rather provides a template for local communities to define their own metric configurations through deliberative processes.

4. Metrics Design and Structural Trade-offs

Operationalizing the theoretical dimensions requires a set of concrete metrics that can be computed automatically or through human evaluation. We propose three classes of metrics: content-based metrics, process-based metrics, and outcome-based metrics. Content-based metrics analyze the generated image for culturally relevant features using pre-trained classifiers that are sensitive to cultural attributes such as clothing, architecture, symbols, and color symbolism. For instance, a metric might measure the presence of culturally appropriate head coverings in images generated for Middle Eastern prompts. Such classifiers must themselves be developed with cross-cultural validation to avoid reifying stereotypes. Process-based metrics evaluate the transparency of the generation pipeline, including whether the model made adjustments based on cultural context tags, whether it issued warnings about potential cultural insensitivity, and whether the user can inspect the influence of specific training examples. Outcome-based metrics assess the real-world impact of the generated content on users, for instance through surveys that capture perceived respect, trust, and harm.

Each metric class involves structural trade-offs that must be managed at the system level. Content-based metrics are computationally efficient and can be integrated into automated auditing pipelines, but they risk oversimplifying cultural variation into fixed categories that may not reflect intra-cultural diversity. Process-based metrics provide deeper insight but require additional runtime overhead and may expose proprietary model internals. Outcome-based metrics are the most ecologically valid but are expensive to collect at scale, subject to self-report biases, and difficult to standardize across languages and cultures. An effective cross-cultural explainability infrastructure will likely combine all three classes, using content-based metrics for large-scale screening, process-based metrics for in-depth audits, and outcome-based metrics for periodic recalibration.

Architecturally, these metrics can be embedded as a separate explainability module that sits between the generative model and the user interface. This module would receive the generated image along with metadata about the user's declared cultural affiliation or inferred cultural context. It would produce a set of explainability scores and, if necessary, flag outputs that fall below predetermined thresholds. The thresholds themselves would be configurable by cultural groups or regulatory bodies. A critical design consideration is that the explainability module must itself be explainable, meaning its decisions should be auditable by independent evaluators. This creates a recursion challenge: the explainability of the metrics requires its

own culturally appropriate explanations. We address this by advocating for a layered approach where the module's logic is made transparent through visual dashboards, narrative summaries, and optional human oversight.

5. Architectural and Governance Implications

Deploying cross-cultural explainability metrics at scale requires significant changes to the architecture of generative AI platforms. Current systems typically treat all users uniformly, with a single global model and a minimal explanation interface. To support cultural adaptability, platforms must implement user segmentation that respects privacy and autonomy. Users should be able to voluntarily provide information about their cultural background or preferred explainability style, either through explicit settings or through behaviorally inferred cues such as language and location. The system must then retrieve a culturally tailored explainability configuration, including the set of metrics, thresholds, and explanation templates.

Governance of such a system raises complex questions about accountability and oversight. Who determines the cultural categories and the associated metric thresholds? How are disagreements between cultural groups resolved when a generated image is simultaneously acceptable to one group and offensive to another? We argue for a polycentric governance model in which multiple independent auditing bodies, representing diverse cultural traditions, certify the explainability metrics and their calibration. These bodies would be analogous to institutional review boards but specialized for AI ethics. The generative platform would be contractually obligated to adhere to the certified configurations for users belonging to each cultural group. Enforcement could be achieved through regulatory licensing, similar to how medical devices require approval for specific populations.

Another architectural implication concerns data governance. The training data for generative models often lacks cultural annotations, and collecting such annotations requires careful engagement with communities to avoid extractive practices. Participatory data collection protocols, as advocated by [12], should be used to co-design culturally relevant taxonomies. The resulting labeled datasets must be stored and managed with appropriate consent and attribution mechanisms. Blockchain or other decentralized ledgers could enable provenance tracking of cultural annotations, ensuring that communities retain control over how their cultural knowledge is used.

6. Case Illustrations and Cross-Domain Comparisons

To ground our framework, we consider illustrative cases from three domains: AI-generated visual content, healthcare AI, and autonomous driving. In visual content generation, consider the prompt "a business meeting." A Western-trained model might output a scene with people shaking hands in a boardroom, whereas in Japan, business meetings often involve bowing and exchanging name cards, and in many Middle Eastern contexts, gender segregation may be expected. A cross-cultural explainability metric would flag the default output as having low representational accuracy for non-Western audiences and would provide an explanation detailing that the training data predominantly featured Western boardroom scenes. The system could then offer an alternative with appropriate adjustments, along with an explanation of the cultural reasoning.

In healthcare AI, diagnostic algorithms often produce recommendations that appear neutral but embed cultural assumptions about health behaviors. For example, an AI system recommending dietary changes might assume a typical American diet. A cross-cultural

explainability metric would evaluate whether the recommendation is accompanied by contextual transparency about its cultural assumptions and whether it aligns with the patient's normative dietary practices. The comparison reveals that explainability in healthcare is more safety-critical and must adhere to medical standards, whereas in visual content, aesthetic and dignity considerations are paramount. In autonomous driving, cultural differences in driving etiquette, such as the meaning of horn use or yielding behavior, affect how explanations for autonomous vehicle decisions should be framed. A system that explains an emergency stop through Western legalistic language may be less intelligible to a driver from a society where informal norms dominate. These cross-domain comparisons demonstrate that while the core dimensions of cross-cultural explainability transfer, the specific metric weights and explanation formats must be domain-specific.

7. Robustness, Fairness, and Sustainability Considerations

Implementing cross-cultural explainability metrics introduces new challenges for system robustness and fairness. Robustness concerns arise because cultural contexts are dynamic and contested. A metric calibrated today may become outdated as cultural norms evolve, or it may be exploited by adversarial users who manipulate their declared cultural affiliation to bypass restrictions. To address this, metrics must be designed to be adaptive, incorporating feedback loops that allow communities to update thresholds and flag erroneous classifications. However, adaptive metrics also risk instability if updates are too frequent or if they are captured by vocal minorities. A governance mechanism that requires supermajority consent for threshold changes can mitigate this risk.

Fairness in cross-cultural explainability metrics is a second-order fairness problem: we must ensure that the metrics themselves do not systematically disadvantage certain cultural groups. For instance, if a metric for representational accuracy uses a Western-trained image classifier to detect cultural features, it may perform poorly for non-Western features, leading to inflated error rates for those groups. This mirrors the problem of algorithmic fairness audits that themselves exhibit bias [13]. Therefore, each metric must be validated independently on culturally diverse test sets, and the validation results must be publicly reported. Another fairness concern is that smaller or less powerful cultural groups may lack the resources to participate in metric design, resulting in invisibility or misrepresentation. Funding mechanisms and capacity building initiatives are needed to support community-led metric development.

Sustainability of cross-cultural explainability infrastructure involves both computational and social dimensions. Continuously running multiple culturally calibrated explanation modules increases inference cost and energy consumption. Techniques such as model compression and selective invocation, where the explainability module is activated only when the model predicts low certainty about cultural alignment, can reduce overhead. Social sustainability requires that the system remains responsive to cultural change without imposing excessive burden on community members. Periodic reviews, perhaps every two years, aligned with major cultural shifts or migration patterns, can balance stability and adaptability. Finally, the knowledge base used to derive metric thresholds must be maintained as a living archive, similar to the concept of dynamic cultural databases proposed in [14].

8. Policy and Deployment Pathways

Regulatory frameworks for AI are gradually incorporating cultural dimensions, but existing proposals remain high-level. The EU AI Act [9] classifies systems based on risk, but cultural

harm is not explicitly recognized as a risk category. We recommend that future regulations include a requirement for cross-cultural explainability auditing for any generative AI system deployed across multiple jurisdictions. This auditing would evaluate whether the system's explainability metrics satisfy minimum standards for representational accuracy, contextual transparency, normative alignment, and user-centered intelligibility for each target cultural context. Auditing bodies should be multidisciplinary, including cultural anthropologists, ethicists, and community representatives.

For deployment pathways, we propose a phased approach. In the first phase, platforms would implement a minimal version of cross-cultural explainability by providing users with a basic cultural context tag associated with each generated image, akin to a content warning but positive. In the second phase, platforms would develop automated metrics for the most widely used cultures, based on existing ethnographic research. In the third phase, a participatory feedback system would allow communities to refine metrics and thresholds. Finally, independent certification bodies would validate the entire process. This phased deployment reduces initial resistance and allows iterative learning.

International cooperation is essential because generative AI systems are global by default. Bilateral and multilateral agreements, such as those under the OECD AI Policy Observatory [8], can establish mutual recognition of certification schemes. However, careful attention must be paid to power asymmetries: wealthier nations should not impose their explainability standards on others. Instead, a pluralistic framework that respects cultural sovereignty while ensuring baseline protections is necessary.

9. Conclusion

Cross-cultural explainability metrics represent a critical frontier for ethical compliance in AI-generated visual content. This paper has argued that existing explainability approaches are insufficient because they assume a culturally homogeneous user base and fail to account for the diverse interpretive frames through which visual content is evaluated. We have proposed a multi-dimensional metric framework that integrates representational accuracy, contextual transparency, normative alignment, and user-centered intelligibility, and we have discussed the structural trade-offs, architectural requirements, governance models, and policy implications of implementing such metrics at scale. The framework is not intended as a final solution but as a starting point for ongoing interdisciplinary dialogue among computer scientists, cultural anthropologists, ethicists, and policymakers. Future work should focus on empirical validation of the proposed metrics in specific cultural contexts, development of open-source tools for community-driven metric design, and longitudinal studies of the impact of explainability interventions on user trust and perceived fairness. As generative AI systems become ever more embedded in daily life, the ability to explain their outputs in culturally resonant ways will be essential for maintaining social license and ensuring that technological progress benefits all of humanity, not just a culturally narrow subset.

References

1. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (pp. 610–623). https://doi.org/10.1145/3442188.3445922

2. Liang, P. P., Wu, C., Morency, L. P., & Salakhutdinov, R. (2023). Towards understanding macro-level biases in visual recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 15712–15721).

3. Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608.

4. Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence, 267, 1–38. https://doi.org/10.1016/j.artint.2018.07.007

5. Hofstede, G. (2001). Culture's consequences: Comparing values, behaviors, institutions, and organizations across nations (2nd ed.). SAGE Publications.

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. Markus, H. R., & Kitayama, S. (1991). Culture and the self: Implications for cognition, emotion, and motivation. Psychological Review, 98(2), 224–253. https://doi.org/10.1037/0033-295X.98.2.224

8. OECD. (2019). OECD principles on artificial intelligence. OECD Publishing. https://www.oecd.org/digital/artificial-intelligence/principles/

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

10. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM Computing Surveys, 54(6), 1–35. https://doi.org/10.1145/3457607

11. Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2021). Datasheets for datasets. Communications of the ACM, 64(12), 86–92. https://doi.org/10.1145/3458723

12. D'Ignazio, C., & Klein, L. F. (2020). Data feminism. MIT Press.

13. Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. In Proceedings of the Conference on Fairness, Accountability, and Transparency (pp. 59–68). https://doi.org/10.1145/3287560.3287598

14. Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

15. Sorensen, T., Jiang, L., Hwang, J. D., Welleck, S., Demberg, V., Durmus, E., ... & Pavlick, E. (2024). A systematic analysis of cultural assumptions in language model training data. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 697–716).

16. Wang, Z., Liu, Y., & Singh, D. (2023). Cultural sensitivity in text-to-image generation: A benchmark and evaluation framework. arXiv preprint arXiv:2308.05492.

17. Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., ... & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (pp. 33–44). https://doi.org/10.1145/3351095.3372873

18. O'Neil, C. (2016). Weapons of math destruction: How big data increases inequality and threatens democracy. Crown.

19. Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. New Media & Society, 20(3), 973–989. https://doi.org/10.1177/1461444816676645

20. Floridi, L., & Cowls, J. (2019). A unified framework of five principles for AI in society. Harvard Data Science Review, 1(1). https://doi.org/10.1162/99608f92.8cd550d1

21. Venkatasubramanian, S., & Alfano, M. (2020). The philosophical basis of algorithmic fairness. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (pp. 1–10). https://doi.org/10.1145/3351095.3372857