AI & Culture Alignment: Interpretation over Measurement.

Eric J. W. Orlowski[*1] Hakim Norhashim[*2]
[*]Equal contribution [1][AI Singapore, The National University of Singapore, 3 Research Link, #02-04 Innovation 4.0, Singapore 117602] [2][AI Singapore, The National University of Singapore, 3 Research Link, #02-04 Innovation 4.0, Singapore 117602]

Correspondence to: Eric J. W. Orlowski ejwo@nus.edu.sg

1. Introduction

As AI systems are increasingly deployed across scientific research, industrial R&D, and public-sector contexts, they operate within culturally dense environments shaped by tacit norms, institutional practices, and cross-cultural collaboration. In these settings, cultural alignment becomes a governance-relevant concern, affecting trust, coordination, and interpretability rather than merely user experience. Yet cultural alignment is commonly operationalised through quantitative proxies and benchmark-based evaluation, treating culture as a measurable object rather than an interpretive phenomenon. This paper, submitted for the subject area "policymaking & governance" at AI4X, argues for a reframing of cultural alignment as an interaction-level, user-situated achievement and examines the implications this has for how cultural alignment in AI models should be evaluated.

2. What Cultural Alignment Ought to be

Prevailing approaches to cultural alignment in AI often treat culture as a representable object: a set of values, facts, or demographic proxies that can be encoded into data and evaluated through simple benchmarks (see, e.g.: Adilazuarda 2024). Whilst convenient, this framing rests on a category error. Culture is not a stable dataset, nor a property that models can meaningfully "possess." It is lived, enacted, and interpreted in context (Causadias 2020). Accordingly, cultural alignment cannot be reduced to conformity with decontextualised proxies or externally asserted ground truths.

We argue that cultural alignment is best understood as an interaction-level achievement. A system is culturally aligned not in the abstract, but for a particular user, in a particular situation, and for a particular purpose. Alignment emerges through situated meaning-making, where users interpret AI outputs within shared social and institutional contexts. Cultural appropriateness is therefore relational and contingent, rather than universally verifiable.

We therefore propose that culturally aligned systems should aim to produce thick outputs: responses that go beyond surface correctness to reflect contextual awareness, social framing, and sensitivity to user intent. Drawing on interpretive social science (Geertz 1973), thickness captures the difference between merely describing an action and rendering its cultural significance. Thick outputs do not claim comprehensive cultural representation; rather, they provide sufficient contextual richness for users to interpret outputs as culturally intelligible within a given interaction.

3. Implications for Evaluation

If cultural alignment is an interaction-level, interpretive achievement rather than a model-level property, prevailing evaluation practices must be reconsidered. Most existing approaches rely on quantitative benchmarks that operationalise culture through proxies and assess alignment via decontextualised correctness-based scoring (Cahyawijaya et al. 2024). While such measures can provide useful signals at scale, they are poorly suited to capturing whether AI outputs are experienced by users as culturally intelligible or appropriate in context.

From this perspective, cultural alignment cannot be evaluated solely against an external, context-independent ground truth. Cultural meaning emerges through users' interpretations, shaped by social position, institutional setting, and intent. Evaluation must therefore attend to how users make sense of AI outputs, not merely whether outputs match predefined cultural labels. In effect, this asks whether a system's outputs are interpreted as plausibly competent within users' own cultural frames.

This does not entail rejecting quantitative evaluation. Quantitative measures remain valuable for identifying patterns, regressions, and system-level behaviours. However, when

used in isolation, they risk mistaking proxy compliance for genuine alignment. Quantitative signals must therefore be supplemented with human-centred evaluation, where users' interpretive judgements are treated as first-order evidence.

4. Evaluation in Practice

Prevailing evaluation approaches, particularly benchmark-based methods, typically standardise evaluation through fixed prompts and response formats. For cultural alignment, this standardisation is often misplaced. If cultural alignment is realised through situated interpretation, then prompting practices, interaction trajectories, and users' interpretive judgements are constitutive of what is being evaluated. Standardising these elements risks reducing cultural alignment to a decontextualised proxy. This therefore gives rise to a methodological tension: evaluation is expected to be standardised to support deployment decisions (Raimondo et al., 2023), yet cultural alignment is inherently context-dependent and cannot be verified in abstraction as a stable, model-level property. We treat this tension as a design constraint on evaluation.

We propose a red teaming–based (Ganguli et al. 2022, Booth et al., 2024) qualitative cultural alignment evaluation methodology that shifts the locus of standardisation away from fixing interactions (prompt and responses), toward conceptual and scope definition and subsequent qualitative data analysis. Evaluation proceeds by first bounding the context under test, including the relevant use cases and user profiles, thereby making explicit the claims of alignment being assessed. Users drawn from the specified context then interact with the system in realistic use cases, with prompts and interaction trajectories left unconstrained. Instances of perceived misalignment are collected and analysed using shared thematic and severity frameworks, producing standardised evaluative outputs suitable for governance and deployment decisions.

Across these levels, Geertzian thickness functions as a shared evaluative expectation: culturally aligned systems should produce outputs that are sufficiently contextually rich to be interpretable by culturally embedded users. Importantly, evaluation under this framework is not oriented toward certifying a model as culturally aligned in general, but toward identifying bounded patterns of misalignment that support iterative improvement, mitigation and informing deployment decisions. Where necessary, comparisons across models can then be made by examining differences in the themes, frequency, and severity of misalignment within the same scoped setting. Where operational requirements call for summary statistics, quantitative measures may be derived from thematic and severity-coded evidence, ensuring that any metrics remain anchored to the evaluative context rather than substituting for it.

5. Conclusion

We have argued that evaluating cultural alignment requires moving beyond proxy-based benchmarks toward approaches that account for how users interpret AI outputs in context. Treating cultural alignment as an interaction-level phenomenon highlights the limits of decontextualised measurement and motivates the use of human-centred, qualitative evaluation methods. By proposing a bounded, red teaming–based evaluation approach grounded in interpretive expectations of output "thickness," this work offers a governance-oriented pathway for assessing cultural alignment without claiming universal cultural representation. As AI systems increasingly mediate scientific and institutional decision-making, such evaluative approaches are essential for responsible deployment and meaningful alignment.

Acknowledgements

Special thanks to Prof. Simon Chesterman, Prof. Jungpil Hahn of AI Singapore / The National University of Singapore for supporting this ongoing project.

Special thanks to Tristan Koh Ly Wey for your previous input, allowing us to take this work many steps forward.

References

[1] Adilazuarda, Muhammad Farid, Sagnik Mukherjee, Pradhyumna Lavania, Siddhant Shivdutt Singh, Alham Fikri Aji, Jacki O'Neill, Ashutosh Modi & Monojit Choudhury. 2024. Towards Measuring and Modeling "Culture" in LLMs: A Survey. Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. *EMNLP 2024*.

[2] Booth, Harold, Murugiah Souppaya, Apostol Vassilev, Michael Ogata, Martin Stanley & Karen Scarfone. 2024. Secure software development

practices for Generative AI and Dual-Use foundation models. *NIST Special Publication 800. U.S. Department of Commerce*. https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-218A.pdf Last accessed: 13 Jan 2026.

[3] Causadias, José M. 2020. What is culture? Systems of people, places, and practices. *Applied Developmental Science* 24(4):310-322.

[4] Cahyawijaya, Samuel et al. 2024. Cendol: Open Instruction-tuned Generative Large Language Models for Indonesian Languages. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers).

[5] Ganguli, Deep, Laine Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, Andy Jones, Sam Bowman, Anna Chen, Tom Conerly, Nova DasSarma, Dawin Drain, Nelson Elhage, Sheer El-Showk, Stanislav Fort, Zac Hatfield-Dodds, Tom Henighan, Danny Hernandez, Tristan Hume, Josh Jacobson, Scott Johnston, Shauna Kravec, Catherine Olsson, Sam Ringer, Eli Tran-Johnson, Dario Amodei, Tom Brown, Nicholas Joseph, Sam McCandlish, Chris Olah, Jared Kaplan & Jack Clark. 2022. Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviours, and Lessons Learned. *ArXiv*. https://arxiv.org/abs/2209.07858.

[6] Geertz, Clifford. 1973. Thick Description: Towards an Interpretative Theory of Culture. https://people.ucsc.edu/~ktellez/geertz1973.pdf. Last accessed: 8 April 2025.

[7] Raimondo, G. M., & L. E. Locascio. 2023. Artificial Intelligence Risk Management Framework (AI RMF 1.0). *National Institute of Standards and Technology*. https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf Last accessed: 13 Jan 2026.