Cultural Incongruencies in Artificial Intelligence

Vinodkumar Prabhakaran
Google Research
San Francisco, US
vinodkpg@google.com

Rida Qadri
Google Research
San Francisco, US
ridaqadri@google.com

Ben Hutchinson
Google Research
Sydney, Australia
benhutch@google.com

Abstract

Artificial intelligence (AI) systems attempt to imitate human behavior. How well they do this imitation is often used to assess their utility and to attribute human-like (or artificial) intelligence to them. However, most work on AI refers to and relies on human intelligence without accounting for the fact that human behavior is inherently shaped by the cultural contexts they are embedded in, the values and beliefs they hold, and the social practices they follow. Additionally, since AI technologies are mostly conceived and developed in just a handful of countries, they embed the cultural values and practices of these countries. Similarly, the data that is used to train the models also fails to equitably represent global cultural diversity. Problems therefore arise when these technologies interact with globally diverse societies and cultures, with different values and interpretive practices. In this position paper, we describe a set of cultural dependencies and incongruencies in the context of AI-based language and vision technologies, and reflect on the possibilities of and potential strategies towards addressing these incongruencies.

1 Introduction

Artificial Intelligence (AI) research includes the demonstration by machines of human-like intelligence and capabilities, measured in terms of how well they match human behavior on certain tasks. This is reflected in current AI frontiers including language understanding and generation, image understanding and creation, knowledge representation and reasoning, and the automatic acquisition (Machine Learning) of capabilities that support these and other domains. However, human behavior is inherently shaped by the cultural contexts humans are embedded in, the values and beliefs they hold, and the social practices they follow. Thus, any AI system mimicking human behaviour will reflect this culturally-shaped understandings of behaviour. However problems can arise when the culture that develops and shapes the AI differs from the globally diverse cultures of the human-AI interaction contexts, due to tensions and misalignments between cultural values and practices. We posit the following two foundational questions: (1) Which aspects of AI systems are dependent on culture, for instance, an AI system's appropriateness for a given ecosystem may depend substantially on the cultures of the humans in that ecosystem?, and (2) What incongruencies and harms emerge when there is a mismatch between an AI system's implicit cultural predispositions and the cultural ecosystem it is used within? In this paper we expand on these questions, and examine the limits of and strategies towards building culturally cognizant AI systems.

2 Cultural incongruencies and associated harms

Cultural dependencies of AI systems are rarely accounted for in current AI research and development work. It is beyond the scope of this short position paper to summarize the myriad existing definitions, across many disciplines, of the term "culture" (for an overview, see, e.g., [22, 20]). First, we focus on cultures created within broader societies demarcated geographically through national and regional

Presented at the First Workshop on Cultures in AI/AI in Culture (non-archival), NeurIPS 2022.

boundaries and not cultures of, e.g., specific organizations. Secondly, we focus on those aspects of culture that exhibit significant variation across human societies, including worldviews, belief systems, and social practices. Due to the central importance of communication and interpretive practices within culture [15, 12], it follows immediately that communicative and interpretive AI technologies, such as NLP and computer vision, have deep cultural dependencies at various levels.

At a high-level, we can distinguish two ways in which culture interacts with AI systems: *in development* and *in use*. The development process of AI systems interface with culture both through the data and resources that capture culturally shaped human behavior, as well as through the cultural norms and values embodied by the developers and researchers themselves. For instance, modern AI systems that are trained or pre-trained on web data may capture various modalities of human behavior, including language use and images, which implicitly bakes in various cultural aspects that then influence downstream applications. Since language and symbols, and ontology and axiology, play a critical role in the development of AI systems—e.g., through "labels" on data, and how "knowledge", "objectivity" [5], "reality/truth" [24], and "system objectives" are constructed—the cultural norms of the AI developers and researchers also pervasively infuse the AI systems [10, 6, 23].

On the other hand, how AI systems are used, whether they perform the tasks they are built for in ways that adhere to culturally shaped expectations, and how they interact with other human behaviours are all culturally dependent. For instance, interpretive tasks are inherently shaped by the culture within which they are embedded in, including not only the cultural-linguistic dependencies [16, 17] of tasks such as inferring emotion, sentiment, offensiveness, but also image and symbol interpretation—including gestures, facial expressions, taboo imagery including pornography and violence, and denotations and connotations of symbols [9, 2, 4, 3]. When the cultural assumptions and norms that are baked into the AI systems during its development are at odds with the cultural norms and expectations of the target cultural ecosystems, we see breakdowns and failures such as cultural misinterpretations or cultural misrepresentations, which we collectively call *cultural incongruencies*. In this section, we present five kinds of harms cultural incongruencies may cause:

- **Cultural barriers**: Not accounting for cultural biases in training data often result in disparate performance of AI systems across different cultural contexts, often disadvantaging cultures that are already historically marginalized. For instance, failing to understand or generate certain languages and dialects may cause NLP-based virtual assistants to perform poorly for users who use those languages or dialects. Similarly, question-answering systems may perform worse on questions related to cultural artifacts from certain cultures, owing both to disparities in training data as well as to gaps in any underlying ontologies and databases. Another example is a computer vision system failing to detect or generate objects, events, or movements that are specific to certain cultures; e.g. a woomera (Australian spear thrower) in a photo, or a description-to-depiction text-to-image system rendering better quality images of cultural artifacts specific to one culture than another.

- **Imposing hegemonic classifications**: The cultural categories of AI developers can become embedded in AI systems and then applied to diverse cultural contexts, imposing epistemic practices that are not endemic to the local cultural context [11]. Such categorizations using the classification schemes of the developers' culture can silence or minimize local cultural perspectives while valorizing the hegemonic culture [1].

- **Safety gaps**: With increasing adoption of AI systems, there are also increasing efforts on ensuring the AI systems are safe and fair [8, 25]. However, these safety guardrails fail if they don't account for the target cultural ecosystems [23]. For instance, content moderation systems meant to detect offensiveness and misinformation may miss culture-specific offensive terms and interpretations allowing toxic or violent speech to propagate for some cultural settings [21, 14]. Pedestrian detection systems trained and tested on Western streets may not be effective in cities in the Global South as rules of mobility e.g. what it means to honk and where is it acceptable to cross a road are created collectively within cultures and differ significantly around the world.

- **Violating cultural values**: Lack of consideration for the cultural context in which an AI system is to be deployed may result in violating the norms that are important to those communities [19]. For instance, a generative language model may produce text that are offensive within certain cultures, even if the language is deemed appropriate at large, e.g., mixing words that are sacred with words that are considered profane. Similarly, a computer vision system may violate cultural norms by producing labels or captions that differ from those preferred by members of that culture.

- **Cultural erasure**: Cultural erasure occurs when knowledge, histories, and identities of a particular people are erased either through omission, trivialization or simplification [26]. [13] describes such erasure as 'symbolic annihilation'; i.e., by not being represented, cultures are annihilated from memory if not physically then metaphorically. Such erasure can happen when technologies homogenize diversity of cultural lives, creating simplified caricatures e.g. a text-to-image model rendering a mosque when prompted to symbolize Islam, not recognizing that Islam is a political, historic, artistic or geographical term not just a religious one. Erasure harm is especially problematic in the context of pre-trained models where such erasure is then also propagated to downstream models.

3 A Research Agenda Towards Culturally Cognizant AI

Having outlined the various incongruencies and associated harms that might arise when AI systems interface with cultural contexts, we now turn to a set of high-level concrete research directions that can begin to mitigate these cultural incongruencies. A perfect culturally competent AI system is not the goal here, as culture is not a static variable that can be easily encoded into a technology, rather a complex and dynamic system that is constantly being transmitted and transformed. Instead, we lay out research questions as openings/opportunities to start a conversation on what a culturally cognizant AI system might look like, is it a desirable goal, and if so, how we may get closer to that state.

How do we identify and measure cultural harms of AI systems: The domain of 'culture' has long been studied by anthropologists and sociologists. Yet, as we have argued in this paper cultural dependencies and incongruencies of AI technology call for a focus of AI/ML researchers on measuring cultural harms of emerging technologies. This task would require: developing conceptualizations and operationalizations of culture without falling into the trap of positivism; creating theoretically rigorous and community driven metrics for measuring complex phenomena like cultural erasure; mapping the incongruencies like the ones we highlighted to on-ground harms for cultural systems. This research question can not be undertaken by AI researchers alone, but requires interdisciplinary collaborations with scholars who have long studied the interplay between culture and technology as well as centering the experiences of communities whose cultural systems we seek to study.

How can we build culturally situated evaluations: In order to ensure that AI system evaluations are culturally situated, it is important to evaluate if the test data reflects the distribution or interpretive practices of the ecosystems in which it is used [18]. Fairness evaluations framed in the West may not readily apply to other socio-cultural contexts [23]. Resources and adversarial probes used for testing may not have cross-cultural coverage; for instance, identity terms (e.g., *African American*) used to test for racial biases in language models are often framed within the US context, but they do not capture the axes of discrimination across the globe. Cultures may also differ in the relative importance they place on different evaluation metrics (e.g., the relative weightings of false positives and false negatives), or of how much they value average-case vs. worst-case behavior. More research is needed to build evaluation paradigms that can effectively incorporate such multi-cultural considerations.

How can diverse perspectives be integrated into the AI pipeline: Many incongruencies we discussed emerge from the narrowness of perspectives represented in the training data and labels, and in the values underlying AI development. To what extent can we diversify these perspectives and what impact would these mitigation strategies have? For instance, human raters form a major part of AI workforce, but they are often deemed interchangeable without accounting for the diverse socio-cultural perspectives they bring to the rating task(s) [7]. Replicating data labeling across the all socio-cultural backgrounds across the globe poses challenges at scale; more research is needed to effectively and efficiently incorporate diverse cultural perspectives in rater pools. Similarly, the AI research and development workforce is also not representative of the global cultures. Technologies emerge from the particular norms and cultures of these developers and institutions. Envisioned AI solutions are influenced by their decisions, such as problem formulation, categorization schemes, and objective functions, to name a few. To what extent, can the AI development processes be opened up to more participatory processes that elevate the voices of more communities as co-creators?

Culture is in of itself a complex phenomena that may not necessarily lend itself to being encoded within technologies, as we imagine them today. Similar conversations have occurred on the possibility of machines encoding fluid social concepts such as identity or gender, and debates exist if these can be understood by machines that are geared towards simplification and objectification. Our call for a research agenda for culturally-cognizant AI is thus aimed towards better understanding the possibilities and limits of encoding cultural cognizance in our AI systems.

Acknowledgements

We would like to thank Courtney Heldreth, Chelsey Fleming, Andrew Smart, and Madeleine Clare Elish for the insightful discussions on this topic which has helped shape this paper.

References

[1] Geoffrey C Bowker and Susan Leigh Star. *Sorting things out: Classification and its consequences*. MIT press, 2000.

[2] Kara Chan, Lyann Li, Sandra Diehl, and Ralf Terlutter. Consumers' response to offensive advertising: a cross cultural study. *International marketing review*, 2007.

[3] Daniel Chandler. *Semiotics: the basics*. Routledge, 2007.

[4] Neil Cohn. *The Visual Language of Comics: Introduction to the Structure and Cognition of Sequential Images.* A&C Black, 2013.

[5] Lorraine Daston and Peter Galison. *Objectivity*. Princeton University Press, 2021.

[6] Emily Denton, Alex Hanna, Razvan Amironesei, Andrew Smart, and Hilary Nicole. On the genealogy of machine learning datasets: A critical history of ImageNet. *Big Data & Society*, 8(2):20539517211035955, 2021.

[7] Mark Díaz, Ian Kivlichan, Rachel Rosen, Dylan Baker, Razvan Amironesei, Vinodkumar Prabhakaran, and Emily Denton. Crowdworksheets: Accounting for individual and collective identities underlying crowdsourced dataset annotation. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 2342–2351, 2022.

[8] Emily Dinan, Gavin Abercrombie, A Stevie Bergman, Shannon Spruit, Dirk Hovy, Y-Lan Boureau, and Verena Rieser. Anticipating safety issues in e2e conversational AI: Framework and tooling. *arXiv preprint arXiv:2107.03451*, 2021.

[9] Paul Ekman. Cross-cultural studies of facial expression. *Darwin and facial expression: A century of research in review*, 169222(1), 1973.

[10] Diana Forsythe. *Studying those who study us: An anthropologist in the world of artificial intelligence*. Stanford University Press, 2001.

[11] Michel Foucault. The order of things: An archaeology of the human sciences, 1989.

[12] Clifford Geertz et al. *The interpretation of cultures*, volume 5019. Basic books, 1973.

[13] George Gerbner. Cultural indicators: The case of violence in television drama. *The Annals of the American Academy of Political and Social Science*, 388:69–81, 1970.

[14] Sayan Ghosh, Dylan Baker, David Jurgens, and Vinodkumar Prabhakaran. Detecting cross-geographic biases in toxicity modeling on social media. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 313–328, 2021.

[15] Edward Hall. *The Silent Language*. Anchor books, 1959.

[16] Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, et al. Challenges and strategies in cross-cultural NLP. *arXiv preprint arXiv:2203.10020*, 2022.

[17] Dirk Hovy and Diyi Yang. The importance of modeling social factors of language: Theory and practice. In *The 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 2021.

[18] Ben Hutchinson, Negar Rostamzadeh, Christina Greer, Katherine Heller, and Vinodkumar Prabhakaran. Evaluation gaps in machine learning practice. In *Proceedings of the ACM Conference on Fairness, Accountability and Transparency (FAccT)*, 2022.

[19] Rebecca L Johnson, Giada Pistilli, Natalia Menédez-González, Leslye Denisse Dias Duran, Enrico Panai, Julija Kalpokiene, and Donald Jay Bertulfo. The ghost in the machine has an american accent: value conflict in gpt-3. *arXiv preprint arXiv:2203.07785*, 2022.

[20] Peter Metcalf. *Anthropology: the basics*. Routledge, 2006.

[21] Vinodkumar Prabhakaran, Zeerak Waseem, Seyi Akiwowo, and Bertie Vidgen. Online abuse and human rights: Woah satellite session at rightscon 2020. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 1–6, 2020.

[22] Nigel Rapport. *Social and cultural anthropology: The key concepts*. Routledge, 2014.

[23] Nithya Sambasivan, Erin Arnesen, Ben Hutchinson, Tulsee Doshi, and Vinodkumar Prabhakaran. Re-imagining algorithmic fairness in India and beyond. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 315–328, New York, NY, USA, 2021. Association for Computing Machinery.

[24] John R Searle. *The construction of social reality*. Simon and Schuster, 1995.

[25] Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Vincent Zhao, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Pranesh Srinivasan, Laichee Man, Kathleen Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Aguera-Arcas, Claire Cui, Marian Croak, Ed Chi, and Quoc Le. LaMDA: Language models for dialog applications, 2022.

[26] Gaye Tuchman. Mass media values. *Society*, 14(1):51–54, 1976.