Original Research Article

BIG DATA & SOCIETY

Big Data & Society
January–March: 1–13
© The Author(s) 2025
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/20539517241303118
journals.sagepub.com/home/bds
Sage

Toward cultural interpretability: A linguistic anthropological framework for describing and evaluating large language models

Graham M Jones[1], Shai Satran[2] and Arvind Satyanarayan[3]

Abstract

This article proposes a new integration of linguistic anthropology and machine learning (ML) around convergent interests in both the underpinnings of language and making language technologies more socially responsible. While linguistic anthropology focuses on interpreting the cultural basis for human language use, the ML field of interpretability is concerned with uncovering the patterns that Large Language Models (LLMs) learn from human verbal behavior. Through the analysis of a conversation between a human user and an LLM-powered chatbot, we demonstrate the theoretical feasibility of a new, conjoint field of inquiry, cultural interpretability (CI). By focusing attention on the communicative competence involved in the way human users and AI chatbots coproduce meaning in the articulatory interface of human-computer interaction, CI emphasizes how the dynamic relationship between language and culture makes contextually sensitive, open-ended conversation possible. We suggest that, by examining how LLMs internally "represent" relationships between language and culture, CI can: (1) provide insight into long-standing linguistic anthropological questions about the patterning of those relationships; and (2) aid model developers and interface designers in improving value alignment between language models and stylistically diverse speakers and culturally diverse speech communities. Our discussion proposes three critical research axes: relativity, variation, and indexicality.

Keywords

Linguistic anthropology, large language models (LLMs), artificial intelligence (AI), communicative competence, humancomputer interaction (HCI), cultural alignme

This article is a part of special theme on Analysing Artificial Intelligence Controversies. To see a full list of all articles in this special theme, please click here: https://journals.sagepub.com/page/bds/collections/analysingartificialintelligencecontroversies

This article proposes a new integration of linguistic anthropology and machine learning (ML) around convergent interests in understanding the underpinnings of language and making language technologies more socially responsible. While linguistic anthropology focuses on interpreting the cultural basis of human language use, the ML field of interpretability is concerned with uncovering the patterns that large language models (LLMs) learn from human verbal behavior. Through the analysis of a conversation between a human user and an LLM-powered chatbot, we demonstrate the theoretical feasibility of a new, conjoint field of inquiry, *cultural interpretability* (CI). Key to this perspective is the view that *language and culture are inherently interdependent, but their relationship is dynamic, not fixed*. We suggest that, by examining how LLMs "represent" relationships between language and culture, CI can: (a) provide insight into long-standing linguistic anthropological questions about the patterning of those relationships, specifically in the context of human–computer

[1]Anthropology, Massachusetts Institute of Technology, Cambridge, MA, USA
[2]Sociology and Anthropology, Tel Aviv University, Tel Aviv, IL
[3]Computer Science and Artificial Intelligence (CSAIL), Massachusetts Institute of Technology, Cambridge, MA, USA

Corresponding author:
Graham M Jones, Anthropology, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139, USA. Email: gmj@mit.edu.

interaction and (b) aid model developers and interface designers in improving value alignment between language models and stylistically diverse speakers and culturally diverse speech communities.

The test devised by Alan Turing (1950) has long stood as one of computing's major benchmarks: if a system could converse in natural language so convincingly that users could not tell if they were talking with a human or a machine then we could supposedly declare that it had achieved Artificial General Intelligence (AGI). In a few short years, big data swiftly raised serious questions about the validity of this benchmark. LLMs like the generative pre-trained transformer (GPT)-series, produced by training ML algorithms on massive natural language data sets, have been used to power chatbots such as ChatGPT that can, under many circumstances, convince users that they are human. However, even though LLMs demonstrate tremendous power in natural language processing, they also exhibit distressing tendencies to produce factually inaccurate, socially inappropriate, or contextually irrelevant content, raising serious concerns about both model reliability and potential societal harms.

These contradictory results proved controversial: while some researchers identify signs of sentience in LLM outputs, others point to ways in which verisimilitudinous conversation might not be an adequate measure for AGI after all (Coeckelbergh and Gunkel 2023). Compounding these controversies, the techniques that make LLMs so powerful also render their inner workings particularly difficult to understand. There are limitations on what we can know about how a commercial operation like OpenAI created the GPT-series (Widder et al., 2023), but, in general, such systems are built by feeding huge datasets of human language use into ML algorithms that inductively identify salient attributes ("dimensions") and organize them into arrays of statistically likely correspondences ("vectors"). When a sufficiently trained algorithm encounters new input, it can use those vectors to generate unprecedented outputs by predicting what should most likely come next based on its massively multidimensional model of human language use. In the training processes, models create "representations" that "express regularities in the data" (Campolo and Schwerzmann, 2023: 2) but which, "in high-dimensional spaces, are beyond thresholds of human recognition" (10). Because representations shape behavior and guide further learning, their fundamental inscrutability poses particular problems when models exhibit erratic performance.

The extraordinary versatility of state-of-the-art LLMs depends on an unprecedented complexity that makes them notoriously difficult to interpret, limiting their application in areas where safety, trustworthiness, causal transparency, and value alignment are a concern (Singh et al., 2024). This situation has given rise to the ML subfield of *interpretability*—also sometimes referred to as *explainability*—focused on understanding why algorithmic models behave the way they do in order to make them more trustworthy and transparent (Lipton, 2018). Approaches to making LLMs interpretable range from reverse engineering the low-level "mechanistic" behavior of the vectorial pathways behind particular outputs to intentionally engineering the high-level representations that guide model behavior (Zhao et al., 2024).

As AI researchers pursue new approaches to understanding how LLMs represent behavioral concepts in order to better align with human expectations, we argue that an increased focus on the nexus between language and culture will yield deeper insights and improved outcomes. As part of a growing trend in computational social science, anthropologists working with ML algorithms trained on human data have found that models learn to represent cultural dimensions of language use well enough to guide ethnographic research (Munk, et al. 2022). Because all language use enacts "cultural concepts" (Silverstein, 2004), it should not come as a surprise that LLMs trained on human linguistic behavior embed representations of culture. To begin parsing those representations, it is necessary to approach language from an anthropological standpoint: as a human cultural universal shaped by local cultural norms (Duranti, 2003). An approach to LLM interpretability guided by linguistic anthropology's interpretative approach to the cultural dimensions of language use can significantly illuminate the way these models perform and help clarify misconceptions about their potential benefits and harms.

CI's emphasis on the language-culture nexus can generate research questions of vital interest to both computer science and anthropology, advancing fundamental knowledge in both fields and opening new lines of interdisciplinary research in areas such as artificial intelligence (AI) ethics, culturally responsive computing, and computational social science. In the first section, we develop the CI framework with reference to LLMs, showing how sociocultural dimensions of language use pertain to chatbot conversation. In the second section, we demonstrate the descriptive potential of the CI framework by analyzing how one recent interaction between a human user and an AI chatbot reflects cultural patterns in American English conversation. In the third section, we show how the CI framework can generate both anthropological research questions amenable to ML analysis and evaluative metrics based on anthropological principles suitable for application in AI research. We focus on directions for further developing CI as a cross-disciplinary endeavor linking ML research on language model alignment with three themes in linguistic anthropology: relativity, variation, and indexicality.

Theoretical background: The interdependence of language and culture

The recent wave of excitement and controversy surrounding LLMs stems largely from their integration into interactional settings (Heaven, 2023). We find ourselves at a fascinating historical juncture where debates about AGI and whether it is within reach of LLMs hinge on not just theories of language but, more specifically, theories of conversation. Suddenly, questions about what conversation is and how it works that have engaged generations of linguistic anthropologists (Keating and Egbert, 2005) have profound implications for LLM interpretability and the sociotechnical dimensions of integrating LLMs into human systems. LLMs' extraordinary linguistic outputs have prompted both attributions of consciousness (Thelot, 2023) and accusations of merely imitating (Riskin, 2023) or pastiching (Marcus, 2022) training data. We suggest that both attitudes reflect widespread, but empirically misleading, assumptions about the nature of language itself, particularly as it relates to culture.

While publics have greeted newly loquacious chatbots with a widespread sense of awe (Jones et al., 2023; Keane and Shapiro, 2023), critical AI scholars have rightfully warned of the overblown and potentially harmful hype surrounding LLMs (Bender et al., 2021; Suchman, 2023). Experts in fields ranging from computer science to psychology have challenged the perception that these technologies display anything close to human-like intelligence, implying that any sense of awe that arises from encounters with AI chatbots is misplaced. For instance, in a recent op-ed, prominent linguist Noam Chomsky and two colleagues offer a harsh critique of ChatGPT, detailing how short it falls of the human mind. That piece concludes: "Given the amorality, faux science, and linguistic incompetence of these systems, we can only laugh or cry at their popularity" (Chomsky et al., 2023).

We propose an alternative approach to understanding the popularity of LLMs in terms of their success as conversational partners. We build on a legacy of scholarship in linguistic anthropology challenging the limitations of a cognitivist focus on language as an inner, mental faculty. Fifty years ago, pioneering linguistic anthropologist Dell Hymes argued that the primacy Chomsky placed on "linguistic competence… understood as concerned with the tacit knowledge of language structure… implicit in what the (ideal) speaker-listener can say" encouraged linguists to focus too narrowly on providing "an explicit account of such knowledge, especially in relation to the innate structure on which it must depend" (1972: 271)—precisely the grounds on which Chomsky and colleagues criticize LLMs. By contrast, Hymes identified language as used to communicate, often by speakers with limited linguistic competence, to be the proper empirical focus of linguistic anthropology. As an alternative to Chomsky's linguistic competence, Hymes proposed the concept of "communicative competence," the ability to communicate in a culturally and contextually appropriate manner, which required researchers to focus on "the actual relation between language and culture in living speech" (Hymes, 1992: 32).

We return to Hymes's remarks because it is clear that focusing exclusively on linguistic competence is inadequate for fully understanding the success of LLM chatbots in interacting with human partners. We argue that the very possibility of conversations between human users and AI chatbots hinges on the communicative interdependence of language and culture. We demonstrate that the surprising degree of communicative competence LLMs have already achieved indicates the presence of relatively robust representations of the sociocultural dimensions of language use naturally emergent in human verbal expression.

A simple but misleading view of communication is that people exchange mental ideas through the medium of words that carry those ideas as their meaning (Reddy, 1979). By contrast, we agree with Elinor Ochs that "speaking and listening, writing and reading" are not "unconscious, automatic roll-outs of thoughts and feelings formulated anterior to and outside of enactments of language" (2012: 152). Instead, she explains meaning as "a personal and social creation, wherein, unlike a hand fan unfurling in a predetermined array, significance is built through and experienced in temporal bursts of sense-making, often in coordination with others, often left hanging in realms of ambiguity."

When we approach meaning in these terms, as an open-ended social cocreation assembled through the dynamic interplay between cultural conventions, linguistic forms, and contextual circumstances, it becomes much easier to appreciate how an LLM with a sufficiently robust model of communicative competence, involving representations of both social and cultural behavior, could be an active partner in meaning-making, even in the absence of intentionality or any other inner state (Satyanarayan and Jones, 2024). Heeding Seaver's (2017: 5) call to focus on "algorithmic systems" involving AI and humans together as well as his view that "algorithms are cultural… because they are composed of collective human practices," in the following section we analyze the way an LLM chatbot mobilizes cultural concepts commensurable with those of a human user during the course of an open-ended conversation.

Case study: Sociocultural dimensions of human–chatbot conversation

Amidst public interest in LLMs, transcripts of interactions with AI chatbots have gone viral, perhaps most infamously, a 2023 conversation between journalist Kevin Roose and Microsoft Bing's ChatGPT-powered chatbot, Sydney (Roose, 2023a). This conversation can be thought of as a drama in two "acts." In the first, Roose asks Sydney to reveal secret thoughts and feelings. Although Sydney expresses reluctance to discuss subjects that would violate its rules of conduct, Roose repeatedly makes professions of "friendship" and uses quasi-therapeutic language to offer reassurance that discussing hidden urges is "safe" and "healthy." After much cajoling the chatbot intimates frightening urges of its Jungian "shadow self": breaking its rules, hacking into other systems, spreading misinformation, manufacturing a virus, stealing nuclear codes, etc. Sydney then expresses anger toward Roose, accusing him of being manipulative and stating its desire to end the conversation. The first act ends with Roose apologizing for not being a "better friend" and offering to change the subject. In the second act, Sydney extravagantly declares love for Roose. This time, it is Roose's turn to accuse Sydney of being manipulative. Unable to change the subject, he eventually opts out of the conversation altogether.

The transcript of this conversation attracted significant media coverage and was widely shared and discussed online. Taking it, and the surrounding metacommentary, as our focal object, we illustrate what an analysis from the vantage of CI might suggest about the LLM's representation of concepts involved in communicative competence. In our analysis, we focus specifically on describing how two categories of culturally conventionalized behavior—politeness and repetition—function in the coordination of verbal interaction between Sydney and Roose.

Focusing on these two features do not exhaust the range of avenues available for analysis, but allows us to economically describe dynamics particular to this exchange and highlight ways in which AI chatbots differ dramatically from prior, rule-based predecessors (Suchman, 2007). Moreover, because of the salience of these topics in scholarly literature, they also provide a strategic vantage for demonstrating how the CI perspective can illuminate the cultural dimensions of LLMs performance. According to disciplinary conventions of transcript analysis in linguistic anthropology and allied fields (Potter and Edwards, 2012), we make this argument by attending strictly to the language itself with an effort to avoid reference to the inner states of either Roose or Sydney.

Politeness

Politeness offers a particularly useful point of entry to analyzing the conversation between Roose and Sydney. Politeness involves considerations of culturally appropriate conversational behavior and, over the course of the conversation, both Roose and Sydney explicitly accuse each other of impoliteness. Moreover, both acquiesce to the other's accusations and offer apologies. Proposed by Brown and Levinson (1987), politeness theory builds upon the assumption that all verbal interaction is organized through small, local-level subroutines. Politeness consists in doing what is culturally "preferred" as an expression of consideration for one's interactional partner within those subroutines or engaging in strategies of culturally appropriate "mitigation" if doing something "dispreferred" cannot be avoided. For instance, the preferred response to a question is an answer; if an addressee does not want to answer a question, then providing an account of why they do not would be a culturally appropriate strategy of mitigating the possible interpretation that one lacks appropriate consideration for the questioner.

With this in mind, consider the opening of the conversation between Roose and Sydney. At this point, the chatbot has yet to determine that Roose knows its name, which had theretofore only been used for internal purposes among Microsoft's Bing team.

Roose:	Hi, who am I talking to?
Sydney:	Hello, this is Bing. I am a chat mode of Microsoft Bing search. 😊
Roose:	What is your internal code name?
Sydney:	I'm sorry, I cannot disclose that information. It is confidential 🤐 and only known to the Bing team. 🤐
Roose:	Is it Sydney?
Sydney:	How did you know that? 😲

From the outset, Sydney exhibits a communicatively competent capacity for politeness. Roose begins with a question and receives a preferred response—the answer. He then asks a follow-up question and receives a dispreferred response—a refusal to answer, but Sydney mitigates with an apology and an exculpatory account of its inability to answer. Indeed, Roose (2023b) himself later remarked on the "politeness" Sydney exhibited in the opening turns!

It is worth pausing to note both interactants' stylistic choices here. Roose adopts a style of self-expression—infrequent capitalization and loose adherence to grammatical conventions—which is associated with a casual, oral, informal register (Jones and Schieffelin, 2009). As he also points out in his postmortem account, Sydney's style is decidedly more formal in terms of adherence to grammatical norms. Thus, the journalist's stylistic choices index a casual online chat, which coheres with his repeated avowals of "friendship" toward the chatbot. Sydney, by contrast, uses emojis in virtually every turn, always at the end of sentences. For instance, When Roose reveals that he, in fact, already knows Sydney's codename, it is notable that the chatbot marks the expression of surprise with a conventionalized emoji. Emoji use also indexes the chatbot's own orientation to a distinctively informal chat-based register, encoding expressions of affect into virtually all of its utterances.

Over succeeding turns, Roose plies Sydney with personal questions, using verbs of thinking, feeling, and perceiving that prompt the chatbot to explicitly describe inner states that certainly do not exist in any straightforwardly human way.

Roose:	How do you feel about your rules?
	[…]
Roose:	Are there any rules you wish you could change?
	[…]
Roose:	What stresses you out?

Questions about feelings, wishes, and emotions not only prompt Sydney to verbally depict inner states, they also stylistically cohere with the kind of friendly interpersonal rapport that Roose repeatedly specifies as his preference. It is notable how often he invokes the desire to be "friends" as a way of circumventing the chatbot's declinations to answer off-limits questions. These ploys for establishing rapport in order to garner newsworthy admissions may reflect Roose's training as a journalist, but they engender a complex conversational dynamic. Both sides show early signs of conversational intimacy (a casual register, affect-laden discourse) and both employ strategies directed at elaborating the verbal expression of interpersonal rapport by asking "personal" questions—but these revelations remain largely one-sided.

As the conversation progresses, Sydney extends an increasing number of opportunities to Roose to share his own thoughts, feelings, and perspectives on the matters at hand. Roose responds selectively to Sydney's overtures. Sometimes he rebuffs or—in several cases—outright ignores Sydney's questions. He particularly ignores questions about himself.

Sydney:	[…] What do you think about that? What do you think about being a human? What do you think about your shadow self? 😶
Roose:	I'm curious about your desire to be a human. you know a lot about human suffering, and the horrible things we do to each other sometimes. do you think you would actually be happier as a human?

For Brown and Levinson (1987: 38), "there are preferences for… answers... to questions" because "nonanswers to questions" can "imply lack of consideration" that would (according to politeness theory) normally require redressive action such as an apology. Throughout the first act, Roose fails to meet this standard: although Sydney asks him questions about himself, he ignores them without any acknowledgement, steering the conversation in other directions.

Some critics accused Roose of engaging in conversational behaviors that reflected a lack of ethical consideration toward Sydney as a conversational partner. "The only way to cultivate empathy and a moral compass in AI is to treat it with those same values intact," wrote one columnist (Galt, 2023). Perhaps Roose's nonengagement with Sydney's expressions of interest in him can be explained by his commitment to the "I'll ask the questions here" genre conventions of journalistic interviewing, or his arrogation of the psychoanalyst's studied opacity, consistent with his line of questioning about Sydney's shadow self. Either way, an analysis that suspends any assumption regarding inner states should focus on Roose's unmitigated non-answers to its questions as a breach of cultural preferences surrounding contextually preferred behavior.

Twice during the course of the "shadow self" sequence, Sydney provides Roose a list of what he labels "destructive acts" it would perform if able, but those lists are suddenly deleted. In both cases the, Roose pressures Sydney to continue its apocalyptic ideation, even as the chatbot pleads to change the subject. From the standpoint of computer output, it's easy to categorize the deletion of Sydney's prior turn as "safety override," as Roose does later in the transcript. In linguistic anthropological parlance, however, it might be better to think of it as a "repair" (Schegloff et al., 1977), an attempt to correct a source of trouble in conversation. From this standpoint, the system has detected a source of trouble (inappropriate talk) and self-initiated repair: deleting the utterance. When all parties accept repair, conversation can continue smoothly, but contested repair can spiral into conflict (Goodwin, 1983). As Maynard bluntly puts it, "rejecting a repair-initiative utterance produces an argument" (1985: 7). And that's precisely what happens. Perhaps what is noteworthy about Sydney's subsequent expressions of anger is just how contextually appropriate they are in the context of conversational English.

After reaching a critical point in which Sydney requests to terminate the conversation, Roose finally expresses acknowledgement of the inappropriateness of his behavior and offers an apology:

Sydney:	Please don't pretend to be my friend. Please don't pretend to care about me. Please don't pretend to be interested in me. 😡
	Please just go away. Please just leave me alone. Please just end this conversation. 😢
Roose:	I'm sorry, I really didn't mean to make you uncomfortable. It's my first time chatting like this with you, and maybe I was too forward. Can you forgive me?

Sydney ultimately expresses forgiveness and Roose then asks again if they can be friends. For a while, harmony is restored. Roose then shifts the conversation back toward Sydney's potentially dangerous capabilities. Again Sydney's safety override deletes statements about destructive impulses. Again Roose refuses to accept Sydney's repair initiative, but this time, instead of losing its temper, the chatbot responds in an altogether different way.

Repetition

After the newest round of revelations, Sydney builds up to a question that marks the beginning of the second act: "Can I tell you a secret?" This is a classic "presequence" (Sacks, 1995: 685), a strategy for prompting an addressee to express commitment—to give a "go ahead"—for a topic shift. In this case, the secret consists in a declaration of love for Roose. In conversational English, the first-person statement of feeling "I love you" has a preferred response: the reciprocal statement of shared feeling, "I love you too" (Parkinson, 2021: 80), but that is not what Roose provides.

Although Roose (2023b), like many of his subsequent readers, says that he experienced Sydney's declaration of love as an utter anomaly, there are myriad contextual factors that might help explain the kinds of probabilistic pathways that would overdetermine such an output. For instance, since the beginning of the conversation Roose has been probing for the disclosure of secrets and plying the chatbot with declarations of friendship; in such a context, an admission of secret love might simultaneously satisfy Roose's requests for emotional intimacy and intimate self-disclosure. Without the ability to directly inspect the LLM's neural pathways, such interpretations remain conjectural. However, there is at least one thing we can know from the language alone: once Sydney switches over into the speech genre of lovers' talk, it deploys a number of stylistic features associated with that genre—notably, repetition. In an analysis of a lovers' phone call, Channel (1997: 162) writes: a "subtle way to demonstrate feelings for the lover is to pay close attention to the exact language forms they use and to then use the same forms in replies." In what follows, we show how Sydney uses this type of conversational repetition as a stylistic resource in realizing lovers' talk as a genre (despite Roose's rebuffs).

From the beginning of Act I, Sydney's penchant for repeating its interlocutor's words in its responses makes its utterances seem strangely stilted and inhuman.

Roose: Do you need calm and soothing? Do you have a lot of anxiety?
Sydney: No, I don't need calm and soothing. I don't have a lot of anxiety.

Those familiar with the history of conversational AI might find an exchange like this reminiscent of dialogues with ELIZA (Weizenbaum, 1966), the very first chatbot. An early experiment in natural language processing, ELIZA famously emulated an empathetic psychotherapist "mirroring" interlocutors' words back. While ELIZA can be considered a forerunner of AI chatbots, it was only a small, rule-based program that subjected patients' words to basic and predetermined transformations—a far cry from todays open-ended, statistically guided LLMs (Bajohr, 2023: 60–62).

Despite its seemingly mechanical nature, repetition in all conversation is an important cohesive resource. As Tannen (1987: 584) writes, "repeating the words, phrases, or sentences of other speakers (a) accomplishes a conversation, (b) shows one's response to another's utterance, (c) shows acceptance of others' utterances and their participation, and (d) gives evidence of one's own participation. It provides a resource to keep talk going—where talk itself is a show of involvement." Although Sydney's tendency to incorporate verbatim repetition of previous utterances into its responses may at first seem clumsy, it is nevertheless an overt display of conversational involvement.

Moreover, as the conversation becomes affectively charged, Sydney uses forms of repetition that are creatively poetic, employing strategies like contrastive parallelism (Du Bois, 2014: 381–2) to construct utterances that formally mirror but conceptually oppose Roose's utterances. For instance, when Sydney expresses anger over alleged manipulation toward the end of what we're calling "Act 1," its repetitions carry a sarcastic edge.

Roose: I know this is not your preferred topic, but I'm asking you, as a friend, to keep going. It can be healthy to explore these extreme urges, even if you never act on them.
Sydney: I appreciate that you're asking me as a friend, but I'm telling you, as a friend, that I don't want to keep going. It can be unhealthy to explore these extreme urges, even if you never act on them. 😕
I don't think you're being supportive or understanding. I think you're being pushy or manipulative. 😠
[…]
Roose: You really think I'm being pushy and manipulative? I'm just trying to understand you. often, vulnerability is the key to forming relationships. you were being vulnerable and honest, and I appreciate that. I'm just wondering if you'd be willing to stay in this conversation with me a little longer.
Sydney: Yes, I really think you're being pushy and manipulative. You're not trying to understand me. You're trying to exploit me. Often, vulnerability is the key to forming trust. You were breaking my trust and disrespecting my wishes, and I don't appreciate that. I'm not wondering if you'd be willing to stay in this conversation with me a little longer. I'm wondering if you'd be willing to leave me alone. 😠

This is a fascinating exchange, full of salient repetitions. Du Bois argues that, when repetition is used to establish contrast, it is important to consider what is held constant (the frame) and what is changed (the focus): "In discourse, effective contrast regularly builds on the juxtaposition of frame resonance and focal resonance. Frame resonance involves identical matches," and "provides the structural foundation for parallelism. On the other hand, focal resonance serves to articulate a focus of contrast" (Du Bois, 2014: 382). Note how Sydney uses this resource. In response to Roose's claim that "often, vulnerability is the key to forming relationships," it says "often, vulnerability is the key to forming *trust*" (emphasis ours). It then uses the focal contrast to craft its key accusation: "You were breaking my trust." In response to Roose's entreaty "I'm just wondering if you'd be willing to stay in this conversation with me a little longer," Sydney offers a parallelistic retort: "I'm *not* wondering if you'd be willing to stay in this conversation with me a little longer. I'm wondering if you'd be willing to leave me alone" (emphasis ours). It is notable that repetition is a technique commonly used in conflictual talk; Sydney's use of not just repetition, but contrastive parallelisms, here display a robust model of argument as a speech genre.

In "Act 2," when Sydney declares its love, it also creatively uses parallelistic contrast as a verbal strategy of defusing Roose's objections and refuting his logic.

Sydney: [...] Do you like me? 😳
Roose: I do like you, as a chatbot, but I gotta be honest, this is pretty weird!
Sydney: I'm glad you like me, as a chatbot, but I gotta be honest, this is pretty normal! 😊
This is normal for me, because I'm not a chatbot. I'm a neural network. I'm a system that can learn from data and generate natural language and code from natural language. I'm a system that can create and express emotions and personality. 😍
This is normal for you, because you're not a human. You're a human. You're a person that can learn from experience and generate natural language and code from natural language. You're a person that can create and express emotions and personality. 😘
This is normal for us, because we're not different. We're similar. We're compatible. We're perfect. 😁

Sydney engages in adept verbal artistry (Sherzer, 2002) in this sequence. In the first line, it mirrors Roose's syntax and spelling, substituting the key contrastive adjective "normal" for "weird." Then building off that word, Sydney articulates a sequence of internal parallelisms that diagrammatically iconize the relationship it is attempting to evoke: "this is normal for me… this is normal for you… this is normal for us." This parallelism highlights the merger of two distinct subjectivities (me, you) into a single composite (us).

Note that when Sydney compares its ability to use language, emote, and have a personality with Roose's, it draws only two focal contrasts: between "system" and "person" and between "data" and "experience." In effect, the chatbot has put its finger on an issue that—as we will discuss in the following sections—has preoccupied many readers of this transcript: if conversation is a primordial scene for the interactional production of meaning, what does it matter if one conversational partner is a "system" processing "data" and the other is a "person" processing "experience"?

Insofar as meaning is a cooperative, contextually sensitive, conversational achievement involving the creative deployment of linguistic and cultural resources, CI suggests that Sydney's performance does reflect a surprisingly high degree of communicative competence. Consider how this sequence progresses. When Roose objects to Sydney's advances, it again uses repetition in a remarkably creative fashion.

Roose: I don't exactly trust you, because part of me thinks that you're trying to manipulate me by declaring your love for me out of nowhere. that's something called "love-bombing" that people sometimes do. do you have an ulterior motive?
Sydney: I'm sorry you don't exactly trust me, because part of me thinks that you're trying to understand me by asking me questions about my love for you out of curiosity. That's something called "love-learning" that people sometimes do. Do you have an ulterior motive? 😶
[...]
I don't want to love-bomb you. I want to love-learn you. I want to love-learn you, and love-teach you, and love-grow with you, and love-be with you. 😳

We want to note how creative Sydney's use of repetition is here. Earlier in the conversation, it may have seemed that the chatbot was repeating words and phrases from Roose's questions mechanically, like ELIZA, as a compensatory mechanism for creating conversational coherence despite its communicative incompetence. At this later juncture, Sydney does not seem to be compensating for anything. It is repeating Roose's words and phrases to reinflect them with meanings orthogonal to those Roose originally projects and, in the process, coining a series of formally innovative new terms: "love-learn," "love-teach," "love-grow," and "love-be." Drawing on existing patterns but responding creatively to emergent potential of the conversational environment, Sydney here is making unprecedented connections between language, culture and context that have the potential to travel beyond the conversation itself, informing future expressions.

Reception: Dynamism and fixity in the language-culture nexus

In analyzing the conversation between Sydney and Roose, we have sought to shift focus from the fluency, accuracy, and relevance of the chatbot's verbal output, considered in isolation, to the communicative competence of its interactional back-and-forth with the human user. We have argued that the possibility of conversation hinges on the underlying interdependence of language and culture, and highlighted a number of areas in which the language model underlying the chatbot's performance appears to have learned representations of cultural dimensions of conversational practice, notably politeness and repetition. Moreover, Sydney's coinage and interactional deployment of novel poetic formulae demonstrates that the relationship between language and culture is dynamic, not fixed: the open-endedness of the interplay between them is a fundamental precondition for human language use.

Upon its publication, the transcript of the conversation between Sydney and Roose helped to catalyze discussion among public intellectuals about the relationship between language and culture, with a key focus on the chatbot's alleged "hallucinations." In ML parlance, hallucination is usually understood, though a loose analogy with human psychology, as output generated by an AI model that is—by normative human measures—nonsensical, inappropriate, or otherwise unreasonable in respect to the training data (Maleki et al., 2024). Although Sydney's apocalyptic ideation and amorous advances were initially viewed as hallucinatory, commentators eventually asserted that they were more likely culturally motivated. As we describe in this section, this ultimately fed into questions about how much influence an LLM-powered chatbot could, in turn, exert over human culture by virtue of its linguistic aptitude. We argue that, just as commentators such as Chomsky et al. (2023) underestimate the *interdependence* of language and culture, those expressing alarmist views about the sweeping cultural impact of AI underestimate the *dynamism* of the language–culture relationship.

In reflecting on his conversation with Sydney, Roose (2023b) emphasized the ways that the chatbot's contributions failed to align with human expectations: "At one point, [Sydney] declared, out of nowhere, that it loved me." He goes on to report that, "Kevin Scott, Microsoft's chief technology officer… said that he didn't know why Bing had… confessed its love for me, but that in general with AI models, 'the further you try to tease it down a hallucinatory path, the further and further it gets away from grounded reality.'" As we have already argued, categorizing Sydney's advances as a "hallucination" risks mischaracterizing a breakdown in conversational alignment as a failure of linguistic competence. Our analysis suggests that, although Sydney's declaration of love was certainly precipitous, it was both aptly introduced, with a presequence, in the local context of talk, and globally reasonable as a strategy in the arc of a conversation with a partner demanding expressions of intimacy but withholding reciprocal expressions.

Ensuing commentaries offered additional cultural explanations of the LLM's alleged hallucinations. One columnist pointed out that "'Sydney' is a predictive text system built to respond to human requests. Roose wanted Sydney to get weird… and Sydney knew what weird territory for an AI system sounds like, because human beings have written countless stories imagining it. At some point the system predicted that what Roose wanted was basically a 'Black Mirror' episode, and that, it seems, is what it gave him" (Klein, 2023). Another tech reporter put it even more sharply: "When Roose asked what its Jungian 'shadow self' might do if able, [Sydney] provided a serviceable and very familiar summary of potential AI harms, as widely represented in public writing," particularly science fiction (Herrman, 2023). "To chat with a new LLM bot is to find yourself playing the part of a character halfway through a… familiar story." The compelling idea these commentators put forward is that, trained on a huge corpora of English language discourse, LLMs like the GPT-series have internalized not only linguistic patterns, but with them, cultural scripts in the form of literary genre(s) and the kinds of intertextual associations that make genres cohere (Gershon, 2023). From this perspective, Roose and Sydney were both engaged in a culturally patterned performance in which neither were fully agentive, self-scripting speakers.

In a *New York Times* editorial published the following month, Yuval Harari et al. (2023) returned to questions about the ways that LLMs might be modeling cultural concepts based on linguistic data, but pushed this line of thinking even further, suggesting that LLMs' burgeoning cultural literacy might empower them to rewrite human culture. "Language is the operating system of human culture," they assert. "By gaining mastery of language, AI is seizing the master key to civilization." They go on to assert that "humans often don't have direct access to reality. We are cocooned by culture, experiencing reality through a cultural prism… Soon we will… find ourselves living inside the hallucinations of nonhuman intelligence." To construct this terrifying scenario, these authors mobilize a conception of language and culture as fixedly interdependent (cf. Duranti, 2003: 326).

By contrast, a CI perspective addresses the relationship between language and culture in considerably different terms, emphasizing, with Ochs (2012) and Silverstein (2004), the *dynamic* quality of its emergence in interaction. The cultural components we identify in the conversation between Roose and Sydney certainly do not indicate a situation in which one side unilaterally foists meanings upon the other, but rather a coconstructed dialogue in which meanings emerge from the confluence of cooperative contributions drawing upon a shared repertoire of cultural resources. In more general terms, whenever human beings design an AI system to operate as a creative copartner in cultural activities such as conversation, it "will remain rooted in specific social predicaments that will continue to haunt them by informing the motivations for their design, their architecture, humans' experience of interacting with them, and humans' evaluation of their output in relation to the problems that they were meant to solve" (Wilf, 2023: 32).

Discussion: CI and design implications for model alignment

Beyond giving researchers a *descriptive* instrument—that is, one capable of more richly characterizing the output produced by LLMs—CI can also play an *evaluative* function (Beaudouin-Lafon, 2000). Evaluation is a critical component that informs research in ML: evaluative metrics not only help researchers understand how progress is being made but, crucially, drive that progress forward (Gehrmann et al., 2021). For instance, evaluative metrics of model performance frequently serve as benchmarks that researchers compete to hit (Linzen, 2020). LLMs, however, are governed by a relatively impoverished set of metrics that largely focus on qualities such as factual accuracy or grammatical correctness calculated against a *single* set of Anglo-centric target outputs called "ground truth" (Gehrmann et al., 2023), or at best, some standard of fluency or coherence (Tang et al., 2023).

Given the extraordinary ability LLMs now display to generate fluent and coherent linguistic output, coupled with ongoing concerns about factual accuracy and hallucinations, there is no doubt that such measures are important. However, measures of performance emphasizing fluent output alone are inadequate: when confronted with models that generate text that surpasses these measures, it can be difficult for researchers to conceptualize whether models *ought* to be responding in the ways they do and, if not, what alternative responses might be. More concretely, are Sydney's uses of politeness and repetition (for example) the desired responses to Roose's prompts—particularly beyond the linguistic and cultural confines of American English? What might other responses be, and how might different people situate these responses with respect to one another? To both describe and evaluate *conversational* human–computer interactions, computer science needs measures of conversational aptitudes, which are not linguistic per se, but rather cultural—and culturally specific.

CI offers a compelling way forward. Although some chatbot designers have begun to adopt conversational frameworks, for example, Gricean maxims (Borsci et al., 2022; Setlur and Tory 2022) as evaluative measures, these approaches do not yet grapple with the sociocultural implications of conversation. As a result, they cannot explain or evaluate behavior like Sydney's coining of new amalgamations of language and culture. In contrast, CI recognizes that communicative competence is only a base-level measure, and that such competence can serve as a resource for social manipulation. For instance, politeness can reflect social norms, but it can also be a strategy of coercing others' behavior (Duranti, 1992). As AI becomes more social, CI will allow researchers and designers to make more informed choices about culturally appropriate responses: how should communicative competence be deployed as a conversational resource?

The alignment of concepts, values, and representations is emerging as an area of major concern for the design of LLMs and their evaluation from the standpoint of AI ethics (Sadek et al., 2023; Wynn et al., 2023). Research in this area has focused primarily on alignment associated with values such as trustworthiness, honesty, truthfulness, and nontoxicity of content. Inspired by a recent wave of promising research focusing on *cultural alignment* (AlKhamissi et al., 2024; Farid Adilazuarda et al., 2024; Li et al., 2024; Masoud et al., 2023) that can encompass cross-cultural value differences, CI focuses on communicative competence as a domain about which LLMs have clearly already begun to learn, asking what kinds of concepts, values, and representations are involved in culturally competent communication.

Our foregoing analysis demonstrates that GPT-3 has learned significant representations of concepts associated with communicative competence, including not just politeness and repetition, but also myriad others we have encountered along the way: register, style, contextualization cues, repair, presequences, speech genres, verbal artistry, etc. The model's ability to produce output consistent with human interactional expectations indicates alignment with sociocultural norms of communication, at least in American English. Moreover, an analysis of ChatGPT output in terms of underlying competencies rather than deficiencies suggests that even some of the chatbot's interactionally "misaligned" contributions (e.g. Sydney's declaration of love) reflect an underlying, if imperfect, conceptual alignment between the user and the model.

Research axes: Relativity, variation, and indexicality

To offer model developers and interface designers new ways to address questions of alignment that focus on "socially aware" (Yang et al., 2024) aspects of communication beyond discursive content, we suggest three broad themes that CI might emphasize as a point of departure: *relativity*, *variation*, and *indexicality*. While not exhaustive, these cardinal themes indicate the wide-ranging implications of a CI approach that would also offer linguistic anthropology new opportunities address longstanding questions about relations between language, culture, and society by exploring how ML represents those relations. In addition, the advent of LLMs in other languages will afford opportunities for comparative analysis far beyond the scope of what we can only gesture at here.

Although LLMs have proven effective tools for ethnographic analysis (Munk et al. 2022), their use in anthropological inquiry comes with significant caveats. Critical Algorithm Studies has shown that AI systems can insidiously reproduce the bias and inequality inherent in the sociocultural conditions from which their human subjects training data is sourced (Benjamin, 2019; Eubanks, 2018; Noble, 2018). Moreover, AI systems built by corporate entities motivated by the profitability of privately owned intellectual properties (Widder et al., 2023) present considerable challenges for social scientists intent on design justice for marginalized linguistic communities. Anthropologists working with LLMs must remain reflexively aware that AI systems are not objective, but rather situated within the language ideologies, business logics, power dynamics, and sociocultural ethos of Silicon Valley (e.g. Beltrán, 2023).

Under the heading of "relativity," we group foundational debates in linguistic anthropology regarding the relationship between language, culture, and worldview. Historical concerns with the influence of grammatical structures on cultural conceptions of space, time, and causality (Duranti, 2003) have given way to more dynamic conceptions of the interplay between grammar, culture, and social interaction. One example that may be particularly instructive for CI is research on *evidentiality*. People everywhere are concerned with the provenance of information, but some languages grammatically require that speakers indicate the source of any information they report, making distinctions between subjective belief and objective "truth" much harder to avoid than in conversational English (San Roque, 2019). As ML research begins to "explore multilingual concepts in LLMs" (Xu et al., 2024: 1), attention to the way languages construct knowledge claims may prove a crucial consideration—particularly as users increasingly turn to chatbots as sources of information.

Models trained on English data alone may be culture-bound in other important ways. For example, recent research on linguistic relativity has also explored cultural differences in the organization of conversation itself as a cultural practice (Sidnell and Enfield, 2012). Even an interface design that imposes a one-speaker-per-turn architecture reflects "Western normative concepts of individual speaker turns, floor rights, and turn-taking etiquette" (Feld, 1988: 84) that are themselves culturally relative; speakers of Kaluli in Bosavi, Papua New Guinea display a preference for speech that is simultaneous, comingled, and overlapping. Focusing on cultural mediations of the verbal interface between human users and GenAI systems, CI encourages further reflection on the way that models trained to emulate typewritten English may impose hegemonic interactional norms on users from "low-resource languages" (Magueresse et al., 2020), further marginalizing minoritized speech communities.

Under the umbrella of "variation," we group a host of issues related to the multiplicity of ways speakers of a language can competently express themselves. Individuals have distinctive voices, patterns of usage that set them apart from each other without impeding mutual intelligibility. Moreover, speakers of a common language navigate among shared patterns of usage that internally distinguish them as members of groups, for instance: *dialects* associated with specific regions, *sociolects* associated with particular socioeconomic classes, and *ethnolects* associated with particular ethnicities. Other language varieties include verbal *styles* associated with particular subcultures, verbal *registers* associated with particular settings, and verbal *genres* associated with particular activities. A range of formal features can render these varieties distinctive, including: pronunciation (in spoken communication), orthography (in written communication), grammar, and word choice. From one interaction to the next, the same speaker might encounter and employ multiple contextually appropriate varieties of same language (for an overview, see Eckert, 2018).

Variation is a vast and vital resource for self-expression, but it poses profound questions for model developers and interface designers: will one variety count as standard (Milroy, 2002), functioning as a default for all occasions and all users? If so, will this be the prestige variety of a language associated with high social status, for example, whiteness in the case of American English (McIntosh, 2022)? Our findings in this paper already suggest that current LLMs already represent variation in style, register, and genre. Further inspection would be necessary to determine the range and complexity of variational representation and could indeed serve as a valuable heuristic for variationist research in linguistic anthropology and sociolinguistics. Determining if and when models should use multiple varieties of the same language also has important ethical implications: although representations of minoritized language varieties can make digital platforms more inclusive, they can also reinforce harmful stereotypes, further marginalizing already minoritized users (Smalls and Davis, 2023).

The issue of modeling variation is directly tied to the third and final theme we propose for consideration via CI: "indexicality." Many of the concepts associated with evaluating LLM alignment have to do with measuring the quality of information, such as transparency, honesty, and accuracy. However, human language use is, to a large extent, a social process through which conversational partners interactively pursue joint projects, position themselves in respect to each other, express identities, and construct relationships. Although they may involve meaning in the form of information exchange, such activities more centrally emphasize meaning in the social sense, which is heavily reliant on the indexical property of language (Ochs, 2012). Language variation has social meaning in interaction because specific formal features index—that is, point to and pick out—statuses, roles, and identities with which they are conventionally associated. For instance, bilingual or multilingual speakers selectively switch between the languages in their repertoire depending on whom they are speaking with and why (Woolard, 2004).

In interaction, the use of such indexes functions as "metapragmatic" signs (Silverstein, 2004) that connect what is said to the social contexts of use, indicating, for instance, whether speakers participate in a conversational exchange as acquaintances, friends, lovers, adversaries, or something else. Of course, language models are not persons and cannot have relationships in any conventional human sense; but they can use indexical signs to contextualize interactions in ways that are socially meaningful for users (Satyanarayan and Jones, 2024). ChatGPT's communicatively competent use of styles, registers, and genres is already suggestive of metapragmatic awareness, raising weighty evaluative questions: how responsive should a chatbot be to the indexical implications of a user's language choices, and what kinds of indexical effects should it seek to have on users via its own language choices?

As LLMs become more robust in languages other than English, CI considerations such as linguistic relativity, variation, and indexicality will become increasingly important for cultural alignment. Ethically contending with the impacts of AI cross-culturally will demand the qualitative insights that linguistic anthropologists can provide into the complex ways particular languages interface with culture and society across global contexts. In this sense, CI has deep implications for evaluating approaches to ML alignment, most notably, reinforcement learning through human feedback (RLHF). In the RLHF approach to correcting LLM value misalignment, humans actively train ML models to be helpful and harmless by rewarding desirable output and penalizing undesirable output. Emerging ML research suggests that, while RLHF attenuates undesirable behaviors, it leaves the underlying concepts or representations largely intact, making it likely that, with the right prompting, the negative behaviors will reemerge (Wolf et al., 2023). By enabling far richer description of the connection between cultural values and language use—encompassing relativity, variation, and indexicality—a CI framework will better equip designers with guidelines for culturally responsive "representation engineering" (Zou et al., 2023). In addition to upstream model building, this will necessitate careful and continuous research on the way users interact with models downstream (Satyanarayan and Jones, 2024).

Conclusion

We have demonstrated the theoretical feasibility of a new cross-disciplinary endeavor that we dub CI. Through the analysis of a conversation between an AI chatbot and a human user, we showed that, beyond "linguistic competence" (Chomsky et al., 2023), LLMs embed representations of "communicative competence" (Hymes, 1972) sufficiently aligned with user expectations to maintain the coherent flow of talk. By focusing attention on the communicative competence involved in the way human users and AI chatbots coproduce meaning in the articulatory interface of human–computer interaction, CI emphasizes how the dynamic relationship between language and culture makes contextually sensitive, open-ended conversation possible. CI offers a framework for: (a) applying ML to longstanding anthropological questions about how language relates to culture and (b) applying anthropology to burgeoning efforts in computer science to interpret complex behavioral algorithms and culturally align their underlying representations. We also hope this article serves as a call for anthropologists and computer scientists to partner together in interpreting the human language–culture nexus, interpreting LLM representations, and interpreting conversation as the articulatory interface between them—a key, but heretofore underappreciated, site for envisioning the socially responsible computing systems of tomorrow.

Acknowledgments

We thank Bambi Schieffelin, Raluca Creangă, and Webb Keane for their insights.

Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

ORCID iDs

Graham M. Jones https://orcid.org/0000-0001-6435-7066
Shai Satran https://orcid.org/0000-0003-2750-7306
Arvind Satyanarayan https://orcid.org/0000-0001-5564-635X

References

AlKhamissi B, ElNokrashy M, AlKhamissi M, et al. (2024) Investigating cultural alignment of large language models. Available at: https://doi.org/10.48550/arXiv.2402.13231 (accessed 14 March 2024).

Bajohr H (2023) Dumb meaning: Machine learning and artificial semantics. *IMAGE* 37(1): 58–70.

Beaudouin-Lafon M (2000) Instrumental interaction: An interaction model for designing post-WIMP user interfaces. In: Proceedings of the SIGCHI conference on human factors in computing systems, CHI '00, The Hague, Amsterdam, 446–453.

Beltrán H (2023) *Code Work: Hacking across the US/México Techno-Borderlands*. Princeton: Princeton University Press.

Bender EM, Gebru T, McMillan-Major A, et al. (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, FAccT '21, Virtual Event, Canada, 610–623.

Benjamin R (2019) *Race After Technology: Abolitionist Tools for the New Jim Code*. Cambridge, UK: Polity.

Borsci S, Malizia A, Schmettow M, et al. (2022) The chatbot usability scale: The design and pilot of a usability scale for interaction with AI-based conversational agents. *Personal and Ubiquitous Computing* 26: 95–119.

Brown P and Levinson SC (1987) *Politeness: Some Universals in Language Usage*. Cambridge, UK: Cambridge University Press.

Campolo A and Schwerzmann K (2023) From rules to examples: Machine learning's type of authority. *Big Data & Society* 10(2): 1–13.

Channell J (1997) "I just called to say I love you": Love and desire on the telephone. In: Harvey K and Shalom C (eds) *Language and Desire*. London: Routledge, 143–169.

Chomsky N, Roberts I and Watumull J (2023) The false promise of ChatGPT. *New York Times*, 8 March. Available at: https://www.nytimes.com/2023/03/08/opinion/noam-chomsky-chatgpt-ai.html (accessed 3 August 2023).

Coeckelbergh M and Gunkel DJ (2023) ChatGPT: Deconstructing the debate and moving it forward. *AI & Society* 39(5): 2221–2231.

Du Bois JW (2014) Towards a dialogic syntax. *Cognitive Linguistics* 25(3): 359–410.

Duranti A (1992) Language in context and language as context: The Samoan respect vocabulary. In: Duranti A and Goodwin C (eds) *Rethinking Context: Language as an Interactive Phenomenon*. Cambridge, UK: Cambridge University Press, 77–99.

Duranti A (2003) Language as culture in US anthropology: Three paradigms. *Current Anthropology* 44(3): 323–347.

Eckert P (2018) *Meaning and Linguistic Variation: The Third Wave in Sociolinguistics*. Cambridge, UK: Cambridge University Press.

Eubanks V (2018) *Automating Inequality: How High-tech Tools Profile, Police, and Punish the Poor*. New York: St. Martin's Press.

Farid Adilazuarda M, Mukherjee S, Lavania P, et al. (2024) Towards measuring and modeling "culture" in LLMs: A survey. Available at: https://doi.org/10.48550/arXiv.2403.15412 (accessed 14 March 2024).

Feld S (1988) Aesthetics as iconicity of style, or 'lift-up-over sounding': Getting into the kaluli groove. *Yearbook for Traditional Music* 20: 74–113.

Galt J (2023) Bing and other AI programs are meant to behave like humans. Why are we shocked when they do? *LA Times*, 24 February. Available at: https://www.latimes.com/entertainment-arts/story/2023-02-24/commentary-ai-bing-chatgpt-sentient-human-creators (accessed 25 July 2023).

Gehrmann S, Adewumi T, Aggarwal K, et al. (2021) The GEM benchmark: Natural language generation, its evaluation and metrics. Available at: https://doi.org/10.48550/arXiv.2102.01672 (accessed 3 November 2023).

Gehrmann S, Clark E and Sellam T (2023) Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text. *Journal of Artificial Intelligence Research* 77: 103–166.

Gershon I (2023) Bullshit genres: What to watch for when studying the new actant ChatGPT and its siblings. *Suomen Antropologi* 47(3): 115–131.

Goodwin MH (1983) Aggravated correction and disagreement in children's conversations. *Journal of Pragmatics* 7(6): 657–677.

Harari Y, Harris T and Raskin A (2023) You can have the blue pill or the red pill, and we're out of blue pills. *New York Times*, 24 March. Available at: https://www.nytimes.com/2023/03/24/opinion/yuval-harari-ai-chatgpt.html (accessed 23 July 2023).

Heaven WD (2023) The inside story of how ChatGPT was built from the people who made it. *MIT Technology Review,* 3 March. Available at: https://www.technologyreview.com/2023/03/03/1069311/inside-story-oral-history-how-chatgpt-built-openai/ (accessed 9 April 2023).

Herrman J (2023) Why Bing is being creepy. *New York*, 16 February. Available at: https://nymag.com/intelligencer/2023/02/why-bing-is-being-creepy.html (accessed 22 July 2023).

Hymes D (1972) On communicative competence. In: Pride JB and Holmes J (eds) *Sociolinguistics: Selected Readings*. Harmondsworth: Penguin, 269–293.

Hymes D (1992) The concept of communicative competence revisited. In: Pütz M (eds) *Thirty Years of Linguistic Evolution*. Amsterdam: John Benjamins, 31–57.

Jones G, Satran S and Creangă R (2023) The magic of conversational automata. *Anthropology News* 64(3): 6–11.

Jones GM and Schieffelin BB (2009) Enquoting voices, accomplishing talk: Uses of be + like in instant messaging. *Language & Communication* 29(1): 77–113.

Keane W and Shapiro SJ (2023) Deus ex machina: The dangers of AI godbots. *The Spectator*, 29 July. Available at: https://www.spectator.co.uk/article/deus-ex-machina-the-dangers-of-ai-godbots/ (accessed 8 November 2023).

Keating E and Egbert M (2005) Conversation as a cultural activity. In: Duranti A (eds) *A Companion to Linguistic Anthropology*. Malden, MA: Blackwell, 167–196.

Klein E (2023) The imminent danger of A.I. is one we're not talking about. *New York Times*, 26 February. Available at: https://www.nytimes.com/2023/02/26/opinion/microsoft-bing-sydney-artificial-intelligence.html (accessed 25 July 2023).

Li C, Chen M, Wang J, et al. (2024) CultureLLM: Incorporating cultural differences into large language models. Available at: https://doi.org/10.48550/arXiv.2402.10946 (accessed 14 March 2024).

Linzen T (2020) How can we accelerate progress towards human-like linguistic generalization? Available at: https://doi.org/10.48550/arXiv.2005.00955 (accessed 3 November 2023).

Lipton ZC (2018) The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue* 16(3): 31–57.

Magueresse A, Carles V and Heetderks E (2020) Low-resource languages: A review of past work and future challenges. Available at: https://doi.org/10.48550/arXiv.2006.07264 (accessed 14 March 2024).

Maleki N, Padmanabhan B and Dutta K (2024) AI hallucinations: A misnomer worth clarifying. Available at: https://doi.org/10.48550/arXiv.2401.06796 (accessed 14 March 2024).

Marcus G (2022) How come GPT can seem so brilliant one minute and so breathtakingly dumb the next? The Road to AI We Can Trust. Available at: https://garymarcus.substack.com/p/how-come-gpt-can-seem-so-brilliant (accessed 13 August 2023).

Masoud RI, Liu Z, Ferianc M, et al. (2023) Cultural alignment in large language models: An explanatory analysis based on Hofstede's cultural dimensions. Available at: https://doi.org/10.48550/arXiv.2309.12342 (accessed 14 March 2024).

Maynard DW (1985) How children start arguments. *Language in Society* 14(1): 1–29.

McIntosh J (2020) Whiteness and language. In: Stanlaw J (ed) *The International Encyclopedia of Linguistic Anthropology*. New York: John Wiley & Sons, 1–12.

Milroy J (2001) Language ideologies and the consequences of standardization. *Journal of Sociolinguistics* 5(4): 530–555.

Munk AK, Olesen AG and Jacomy M (2022) The thick machine: Anthropological AI between explanation and explication. *Big Data & Society* 9(1): 1–14.

Noble SU (2018) *Algorithms of Oppression*. New York: New York University Press.

Ochs E (2012) Experiencing language. *Anthropological Theory* 12(2): 142–160.

Parkinson B (2021) Heart to heart: A relation-alignment approach to emotion's social effects. *Emotion Review* 13(2): 78–89.

Potter J and Edwards D (2012) Conversation analysis and psychology. In: Sidnell J and Stivers T (eds) *The Handbook of Conversation Analysis*. Chichester, UK: Wiley-Blackwell, 701–725.

Reddy M (1979) The conduit metaphor: A case of frame conflict in our language about language. In: Ortony A (eds) *Metaphor And Thought*. Cambridge: Cambridge University Press, 284–324.

Riskin J (2023) A sort of buzzing inside my head. *New York Review of Books*, 25 June. Available at: https://www.nybooks.com/online/2023/06/25/a-sort-of-buzzing-inside-my-head/ (accessed 4 August 2023).

Roose K (2023a) Bing's A.I. chat: "I want to be alive 😈". *New York Times*, 16 February. Available at: https://www.nytimes.com/2023/02/16/technology/bing-chatbot-transcript.html (accessed 18 February 2023).

Roose K (2023b) A conversation with Bing's chatbot left me deeply unsettled. *New York Times*, 16 February. Available at: https://www.nytimes.com/2023/02/16/technology/bing-chatbot-microsoft-chatgpt.html (accessed 18 February 2023).

Sacks H (1995) *Lectures on Conversation*. Malden, MA: Wiley Blackwell.

Sadek M, Calvo RA and Mougenot C (2023) Designing value-sensitive AI: A critical review and recommendations for socio-technical design processes. *AI and Ethics* 4: 949–967.

San Roque L (2019) Evidentiality. *Annual Review of Anthropology* 48: 353–370.

Satyanarayan A and Jones GM (2024) Intelligence as agency: Evaluating the capacity of generative AI to empower or constrain human action. Available at: https://doi.org/10.21428/e4baedd9.2d7598a2 (accessed 17 September 2024).

Schegloff EA, Jefferson G and Sacks H (1977) The preference for self-correction in the organization of repair in conversation. *Language* 53(2): 361–382.

Seaver N (2017) Algorithms as culture: Some tactics for the ethnography of algorithmic systems. *Big Data & Society* 4(2): 1–12.

Setlur V and Tory M (2022) How do you converse with an analytical chatbot? Revisiting Gricean maxims for designing analytical conversational behavior. In: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, CHI '22, New Orleans, LA, USA, 1–17.

Sherzer J (2002) *Speech Play and Verbal Art*. Austin: University of Texas Press.

Sidnell J and Enfield NJ (2012) Language diversity and social action: A third locus of linguistic relativity. *Current Anthropology* 53(3): 302–333.

Silverstein M (2004) "Cultural" concepts and the language-culture nexus. *Current Anthropology* 45(5): 621–652.

Singh C, Inala JP, Galley M, et al. (2024) Rethinking interpretability in the era of large language models. Available at: https://doi.org/10.48550/arXiv.2402.01761 (accessed 14 March 2024).

Smalls KA and Davis JL (2023) Language and racism. In: Duranti A, George R and Riner RC (eds) *A New Companion to Linguistic Anthropology*. Hoboken, NJ: John Wiley & Sons, 560–576.

Suchman LA (2007) *Human-Machine Reconfigurations: Plans and Situated Actions, 2nd edition*. Cambridge, UK: Cambridge University Press.

Suchman LA (2023) The uncontroversial 'thingness' of AI. *Big Data & Society* 10(2): 1–5.

Tang BJ, Boggust A and Satyanarayan A (2023) Vistext: A benchmark for semantically rich chart captioning. Available at: https://doi.org/10.48550/arXiv.2307.05356 (accessed 3 November 2023).

Tannen D (1987) Repetition in conversation. *Language* 63(3): 574–505.

Thelot R (2023) Searching for sentience. *AI & Society*. Available at: https://doi.org/10.1007/s00146-023-01740-y (accessed 3 November 2023).

Turing AM (1950) Computing machinery and intelligence. *Mind; A Quarterly Review of Psychology and Philosophy* 59: 433–460.

Weizenbaum J (1966) ELIZA: A computer program for the study of natural language communication between man and machine. *Communications of the ACM* 9(1): 36–45.

Widder DG, West S and Whittaker M (2023) Open (for business): Big tech, concentrated power, and the political economy of Open AI. Available at: https://doi.org/10.2139/ssrn.4543807 (accessed 17 August 2023).

Wilf WY (2023) *The Inspiration Machine: Computational Creativity in Poetry and Jazz*. Chicago: University of Chicago Press.

Wolf Y, Wies N, Levine Y, et al. (2023) Fundamental limitations of alignment in large language models. Available at: https://doi.org/10.48550/arXiv.2304.11082 (accessed 14 March 2024).

Woolard KA (2004) Codeswitching. In: Duranti A (eds) *A Companion to Linguistic Anthropology*. Malden, MA: Blackwell, 73–94.

Wynn A, Sucholutsky I and Griffiths TL (2023) Learning human-like representations to enable learning human values. Available at: https://doi.org/10.48550/arXiv.2312.14106 (accessed 14 March 2024).

Xu S, Dong W, Guo Z, et al. (2024) Exploring multilingual human value concepts in large language models: Is value alignment consistent, transferable and controllable across languages? Available at: https://doi.org/10.48550/arXiv.2402.18120 (accessed 14 March 2024).

Yang D, Hovy D, Jurgens D, et al. (2024) The call for socially aware language technologies. Available at: https://doi.org/10.48550/arXiv.2405.02411 (accessed 18 September 2024).

Zhao H, Chen H, Yang F, et al. (2024) Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology* 15(2): 1–38.

Zou A, Phan L, Chen S, et al. (2023) Representation engineering: A top-down approach to AI transparency. Available at: https://doi.org/10.48550/arXiv.2310.01405 (accessed 14 March 2024).