Last Words

Socially Aware Language Technologies: Perspectives and Practices

Diyi Yang[1], Dirk Hovy[2], David Jurgens[3], and Barbara Plank[4]

[1] Stanford University
 diyiy@stanford.edu

[2] Bocconi University
 dirk.hovy@unibocconi.it

[3] University of Michigan
 jurgens@umich.edu

[4] LMU Munich
 bplank@gmail.com

*Language technologies have advanced substantially, particularly with the introduction of large language models. However, these advancements can exacerbate several issues that models have traditionally faced, including bias, evaluation, and risk. In this perspective piece, we argue that many of these issues share a common core: a lack of awareness of the social factors, interactions, and implications of the social environment in which NLP operates. We call this **social awareness**. While NLP is improving at addressing linguistic issues, there has been relatively limited progress in incorporating social awareness into models to work in all situations for all users. Integrating social awareness into NLP will improve the naturalness, usefulness, and safety of applications while also opening up new applications. Today, we are only at the start of a new, important era in the field.*

1. Introduction

Natural language processing (NLP) has made significant advances in recent years, thanks in part to the introduction of large pretrained language models (LLMs) based on Transformers (Brown et al. 2020). As a result, performance on various NLP tasks has significantly improved, including machine translation, sentiment analysis, and conversational agents, to name but a few. NLP models appear to perform these tasks as well as, if not better than, humans (Tedeschi et al. 2023). On the other hand, an increasing

number of problems and flaws with these models have been identified, which mean that NLP is working unevenly across users and situations. Some of these issues include bias (Bolukbasi et al. 2016; Vida, Damken, and Lauscher 2024), toxicity (Gehman et al. 2020), trust (Litschko et al. 2023), and fairness (Hovy and Spruit 2016; Blodgett et al. 2020; Shah, Schwartz, and Hovy 2020; ElSherief et al. 2021). For example, even basic components such as word embeddings (representing words in a mathematical space) can inadvertently capture and reinforce biases in training data, perpetuating stereotypes and inequalities (Bolukbasi et al. 2016; Gonen and Goldberg 2019; Ryan, Held, and Yang 2024). Machine translation systems can produce translations with unintended biases or inaccuracies (Vanmassenhove, Hardmeier, and Way 2018; Hovy, Bianchi, and Fornaciari 2020), potentially exacerbating cultural and societal misunderstandings (Bird and Yibarbuk 2024). These concerns are exacerbated in widely used models such as LLMs (e.g., Vida, Damken, and Lauscher 2024; Kantharuban et al. 2024; Wilson and Caliskan 2024). All these aspects apply not only to English but also to the 7,000 languages available (Joshi et al. 2020), adding complexity to the problem. Consequently, NLP "works" only for a subset of situations and people that use language technology (Held et al. 2023).

We argue that many of these issues confronting modern NLP have a common core. They are caused by a failure to consider language (technologies) in a social context, that is, in relation to social environments. We refer to these issues as **social awareness**, which refers to a system's awareness of social factors, contexts, and dynamics, as well as their implications for the broader social environment. Social awareness in models is currently undervalued. Traditional NLP models prioritized syntax, grammar, and lexicon, and their modern counterparts are direct descendants. However, they have not significantly progressed in understanding sociocultural context and social interactions. In the linguistic terms of de Saussure, NLP has been mainly concerned with the abstract patterns of language (langue) without paying much attention to the concrete individual use of it (parole). Operationalizing and integrating these complexities into today's LLMs is a significant challenge. However, we argue that addressing this issue is necessary to advance NLP. For example, the simple act of turn-taking in a dialogue, namely, knowing whether, when, and how to respond, requires a certain level of social awareness that LLMs currently lack (Ivey et al. 2024). Without it, conversations can be stilted and come across as rude.

Social awareness is ultimately integral to all modalities of AI, not just NLP, for example, vision (Fathi, Hodgins, and Rehg 2012) and robotics (Breazeal 2003), to name but a few examples. Social awareness governs the dynamics of human–human and human–AI interactions. Language is an essential tool in these processes for people to achieve a wide range of goals. NLP's potential insights and applications will inevitably be limited if it does not consider individual interactions, the context in which language is spoken, and the specific goals it should achieve. Knowing such goals or capabilities allows users to gain more trust in NLP systems (Litschko et al. 2023).

Language is deeply intertwined with human society and culture, making it much more than just words and grammar. By modeling the social factors that influence language use, our models can broaden their scope and depth by better understanding and connecting with people (Hovy and Yang 2021; Hershcovich et al. 2022; Pawar et al. 2024). The goal of this position piece is not simply to propose novel research directions, but to offer a comprehensive discussion of perspectives and practices about socially aware language technologies. It is intended for NLP researchers and practitioners interested in the societal impacts of language technologies, as well as human–computer interaction scholars and social scientists studying their influence on humans and society.

2. What Are Socially Aware Language Technologies?

We define **socially aware language technologies** as *the study and development of language technologies from a social perspective*, allowing NLP systems to understand and respond to social signals expressed in language and broader physical and social environments. Socially aware systems can recognize social aspects and process socially driven meanings and implications behind language in the same way that humans do. In other words, a socially aware system exhibits emotional intelligence, cultural competence, and perspective-taking abilities, as discussed next, to ensure that advances in NLP are technologically sound and socially conscious.

Prior work such as Pentland (2005) defines **socially aware computation** as systems that understand social signaling and context, and further argues that focusing on such dimensions can enhance collective decision-making and keep users informed. The psychologist Daniel Goleman defines **emotional intelligence** as four subsets: self-awareness, self-management, social awareness, and relationship management (Goleman 2006). Emotional intelligence requires understanding and empathizing with others' emotions, which is related to Theory of Mind (Tomasello 2014; Premack and Woodruff 1978). The emphasis on **social awareness** in NLP means creating tasks, models, and evaluations that consider social factors (Hovy and Yang 2021) first (as illustrated in the inner two circles in Figure 1), and then go further by including the full social context and the social dynamics communicated through language. Researchers and practitioners need to become aware of these social aspects to design socially aware NLP systems.

Note that while social awareness includes aspects of cultural and personal identity, it does not require us to take a moral stance, and we do not prescribe *what* perspectives models will have to take (and likely they will differ substantially among different countries and languages). In complex contexts where diverse values and priorities can lead to conflicting outcomes (Sorensen et al. 2024), what benefits one group may harm another. With socially aware language technologies, we can gain a better understanding of such complex situations, as it advocates incorporating context and social factors such
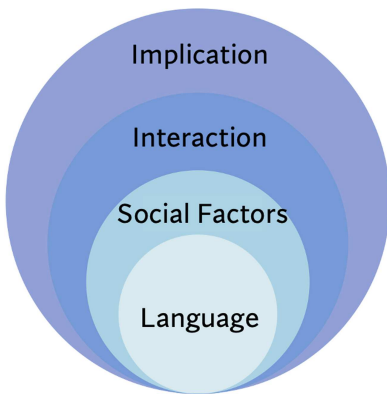
Figure 1
Conceptual structure of socially aware language technologies: social factors, interaction, and implication. This is *not an exclusive partition*, but one way to understand the scope of social awareness.

as culture into the design process, so that resulting technologies reflect the diverse viewpoints rather than prompting dominant ones.

We argue that developing socially aware language technologies must prioritize three key aspects: *social factors* (Section 2.1), *interaction* (Section 2.2), and *implication* (Section 2.3). Figure 1 shows that social awareness is often present near the design of the tasks and algorithms (social factors). Social awareness also plays a central role in the middle-ground of interactions and activities that humans have with NLP systems (interaction) and the outer layer of impact (implication) that NLP systems may have on people and society. By putting a strong emphasis on social factors, interaction, and implications, we hope that socially aware language technologies can facilitate better communication and align with human preferences and societal values.

2.1 Social Factors

We define **social factors** as a wide range of social aspects that shape the way we understand language use, including but not limited to who is the speaker, who is the receiver, what is their social relation, in what context, guided by what kinds of social norms, culture, and ideology, and for what communicative goals (see Hovy and Yang 2021). One can also greatly enrich such a list by incorporating insights and understandings around social factors from psychology, sociology, and other disciplines. There is an increasing trend in doing so in the NLP community, and recent work has highlighted the importance and complexity of evaluating social attitudes, opinions, and values embedded within LLMs (Ma et al. 2024). Social factors, in particular, can motivate socially informed tasks by adding objective functions and tasks. Operationalizing social phenomena will expand the current pool of tasks to reflect users' needs better, resulting in increased user trust. Social knowledge can enhance existing representations in models (Nguyen, Rosseel, and Grieve 2021) and impact opinions and steerability of LLMs (Wright et al. 2024). Social signals may offer alternative supervision for representation learning and next-word prediction. Current models have internal representations of social factors but do not seem to actively draw on them (Lauscher et al. 2022). With social awareness integrated into the pipeline, the outcome can have a social impact, not just on the typical task evaluation metric but also on people.

2.2 Interaction

Interaction refers to the social exchanges and activities between humans and NLP systems, including relational, organizational, and cultural norms that govern inter-personal communication, as well as the evolving contexts in which these systems operate. Social science theories, such as those based on social influence and social norms, define critical dimensions of social interaction, providing insight into how humans interact and behave. Such a perspective posits that language is not an iso-lated construct but emerges as a product of social exchange and communication, aligning closely with the interactionism paradigm in sociology (Snyder and Ickes 1985). Social norms govern social behavior and are defined as groups' shared stan-dards of acceptable behavior. Integrating social norms into language adds expres-sivity beyond vocabulary and grammar. The work of Lapinski and Rimal (2005) highlights the nuanced interplay between social norms and language, demonstrat-ing that linguistic expressions frequently serve as vehicles for the expression and reinforcement of these norms. These aspects provide a rich and multifaceted foun-dation for our explorations of the complicated space of *social interaction*, including

social exchange between individuals, other people in the context, and other activities surrounding the context. Language use, like self-perception (Cooley 1902), is influenced by others' perceptions of the language, especially as people interact with LLMs. Socio-technical NLP systems are part of a social interaction ecosystem in which users, developers, and stakeholders collaborate to develop, deploy, and use these technologies. Many factors and social phenomena are revealed in social interactions, such as power dynamics (Prabhakaran, John, and Seligmann 2013), trust (Litschko et al. 2023), and user expectations (Dhuliawala et al. 2023). NLP system design must consider how social interactions shape user experiences (Jakesch et al. 2023; Liu et al. 2022) and impact the technology's adoption and effectiveness.

2.3 Implication

Implication refers to the impact of an NLP system on society, including both positive and negative effects. Understanding the social implications of NLP is crucial for responsible development and sustainable use. This process involves assessing biases and stereotypes (Dev et al. 2022), considering how systems affect global populations, not just those in the global north (Song et al. 2023; Ranathunga and de Silva 2022), investigating misinformation and dual use of NLP systems, examining concerns about job displacement (Eloundou et al. 2023) and human–LLM alignment on other factors such as, for example, creativity (Spangher et al. 2024) and storytelling (Tian et al. 2024). Understanding social implications can also inspire model design, such as developing models that consider the implications of their outputs. For example, work in prompt safety can be viewed as an initial step towards imbuing models with a sense of what responses have harmful social implications (e.g., Bianchi et al. 2023). As a society, we first have to understand these social implications to fully use NLP and mitigate its harmful effects.

3. Situating Socially Aware Language Technologies

As with many other things, social awareness is best identified by its absence. Without social awareness, NLP technology will disregard social or cultural taboos, fail to consider personalized aspects of language applications, use language that the target audience cannot understand (due to age, education level, or other factors), or respond inappropriately or hurtfully (e.g., telling a suicidal user to kill themselves [Dinan et al. 2022]) or in non-natural ways. Socially aware language technologies are related to many emerging topics, and as more work is being done or needs to be done around social and language technologies, there is a crucial need to differentiate socially aware language technologies from other approaches.

3.1 Differentiation

Socially aware NLP differs from *personalization* because it aims to incorporate a broader context of language use, such as larger social and cultural groups. In contrast, personalization focuses more on the individual for a customized user experience (e.g., Flek 2020). The concepts of socially aware NLP and "NLP in a social context" are related but not the same. Using NLP techniques to analyze and understand language use in social settings such as online communities, political discourse, and public opinion is called NLP in a social context or computational sociolinguistics (Nguyen et al. 2016), sometimes through the text lens of Computational Social Science (CSS). CSS often develops NLP

models to uncover patterns and trends in text to answer questions in the social sciences. Another related concept is the **Theory of Mind** (ToM; Grant, Nematzadeh, and Griffiths 2017; Le, Boureau, and Nickel 2019; Sap et al. 2022), which refers to the ability of models to reason about the mental state of others (e.g., intents, emotions, or beliefs). While ToM and socially aware NLP aim to improve models' ability to interact with humans more easily and socially appropriately, ToM differs from socially aware NLP by focusing on inferring others' mental states through attributing mental states and understanding intentions. *Human-centered NLP* and socially aware NLP both emphasize how to make NLP aware of human factors and align with real-world needs, including ethical considerations and inclusiveness of languages and cultures (Soni et al. 2024). Human-centered NLP, on the other hand, focuses on user-centered design to create systems tailored to user needs and is frequently based on iterative design, usability testing, and human-in-the-loop approaches to improve human-system interactions. Similar comparisons apply to the difference between *human-like* and *social awareness*. Human-like AI aims to mimic human-to-human interactions by making them natural and familiar to humans. At the same time, social awareness *further* encourages appropriate and considerate interactions with the social environment.

4. Building Socially Aware Language Technologies

While NLP has recognized the importance of social language and begun to develop models capable of interacting socially, there are still significant gaps. Here, we motivate the key considerations and strategic goals for closing these gaps.

4.1 Considerations for Socially Aware NLP

Just as language grounding models benefit from diverse imagery to map to language, socially aware models need exposure to massive amounts of diversity—persons, values, relationships, and so forth—in order to learn effective representations of social information and learn to reason about it. Humans naturally form categories to describe stereotypical social entities (e.g., persons, groups, relationships), which help guide expectations for behavior and communication in the absence of more explicit information (Blair, Ma, and Lenton 2001; Rhodes and Baron 2019). Although crude and potentially harmful, such stereotypes can serve as effective priors that simplify social processing *but* are only an initial starting point in a representation of others and one that is updated through interaction—bringing a richness and nuance to how we think of and relate to each other. Given the complexity of social awareness, we argue that models need sufficient diversity to effectively learn these categories so that they can easily adapt and reason about the social cues seen in different settings (Yin et al. 2021; Sharma et al. 2023; Wang et al. 2024). Thus, we argue for the following considerations: social awareness and social reasoning must (C1) be grounded in socially diverse data, with fluid representations that move from the categorical to the bespoke; (C2) extend to interactive contexts, diverse cultural settings, and non-text modalities; (C3) be adaptable to different contexts and be capable of adapting during interaction; and (C4) be interpretable or explainable in how social awareness is used to reason.

4.2 LLMs Are Not Socially Aware Language Technologies Yet

The advanced language capabilities of LLMs have opened up many new interactive and seemingly social applications. However, we argue that LLMs are not yet socially aware

and that we need new goals and measurements to gauge progress and move beyond traditional NLP benchmarks for social tasks (e.g., Choi et al. 2023; Ziems et al. 2023).

(1) *Operationalization and measurement of social awareness.* Many recent studies have started to quantify social awareness (Rathje et al. 2024), like measuring social relations (Iyyer et al. 2016; Choi et al. 2021) or recognizing inappropriate content (Kumar, AbuHashem, and Durumeric 2024). LLMs are shown to struggle with these social signals (Ziems et al. 2023), calling for new algorithms and systems to deal with them. It becomes increasingly important to operationalize different aspects of social awareness based on theories and insights from social science in order to determine whether and to what extent LLMs have exhibited social awareness. Further, such evaluations must go beyond static benchmarks or multiple-choice questions to operate in an interactive way.

(2) *Behavioral expectations from social science theories*. Experimental work in the social sciences such as psychology and behavioral economics has generated clear expectations for a variety of human behaviors, such as trust (Evans and Krueger 2009) and risk aversion (Dohmen et al. 2005), that depend on social awareness. When prompted with similar settings and information, these insights can serve as references for external behaviors demonstrating that LLMs are accurately reasoning about social awareness in a human-like manner (Park et al. 2024). Such experimental measurement allows us to test whether models recognize the social factors in play and to interact accordingly.

(3) *Inference of social context and use in reasoning*. Humans learn to recognize social cues as they mature and reason about this information (Thompson 2007; Sher, Koenig, and Rustichini 2014). Although LLMs are increasingly capable of complex reasoning tasks (Huang and Chang 2023), they are still only beginning to learn to recognize social information through interaction (e.g., Zhang et al. 2023) and to be able to explicitly incorporate and explain how this information influences their reasoning (Gandhi et al. 2024). As one example domain, many social games involve reasoning not only about the game state but players' mental states and the social implications of certain actions (Colman 2003). Games can provide new domains for assessing models' abilities to learn social factors from cues across turns and reason about other players.

(4) *Deployment of socially aware behavior in practical applications*. Technologies that use social factors and implications in real-world applications provide rich ground for assessing progress in social awareness. Recent work has targeted applications such as therapy and coaching (Suh, Althoff, and Torous 2024), inclusive technologies for providing access to services for people with disabilities (Guo et al. 2020), and language technologies for positive impact (Jin et al. 2021). However, clear gaps exist between human social behavior in these applications and the capabilities of LLMs.

(5) *Understanding how socially aware language technologies affect people and society*. With the increase of LLM-empowered applications, it becomes critical to understand these broad implications, which include how LLM-empowered applications affect how people communicate and interact with each other (Liu et al. 2022), reinforce stereotypes or biases (Dev et al. 2022), and affect public trust, education, and the labor market (Eloundou et al. 2023), as well as how they inform policy and regulation.

Taken all together, there is a critical need for a sub-field of *socially aware language technologies* due to the increasing work on social and language technologies. Within this new subfield, we must ensure that language processing advances are technologically sophisticated and socially conscious. A unified subfield focused on this goal would allow researchers to systematically address the challenges of embedding social intelligence into language models, allowing for more precise communication among scientists, policymakers, and the general public. Recognizing *socially aware language technologies*

is a strategic step towards a future where language technology responsibly interacts with human society.

5. Historical View of Socially Aware Language Technologies

Early AI was conceived in a much more holistic manner than the fragmented space that exists today. Its goal was to produce *human-like* behavior, which required a tight coupling of different aspects and disciplines. That goal assumed social awareness, even if not explicitly stated (Turing 1950; McCarthy et al. 2006). Moravec's paradox (Moravec 1988), often summarized pithily as, "In AI, easy things are hard, and hard things are easy" (Pinker 2003), has singled out *social awareness* and *motion* as the main areas where AI models have difficulty matching human performance even on simple tasks (while outperforming humans on tasks that require patience or logic). Over time, AI specialized into subfields, which shifted their focus to easier-to-solve tasks. Those were typically information or logic-based and did not require social awareness. As a result, NLP has spent a long time focusing on information-rich linguistic analysis tasks like parsing. Recent research has focused on language's social, cultural, and demographic aspects (Hovy and Yang 2021; Dev et al. 2023).

LLMs' strong performance on various language understanding tasks may create the superficial impression that these models are now socially aware. However, many of the tasks they excel at are language-only problems that do not necessitate social awareness. Furthermore, tasks designed to demonstrate social, psychological, or emotional aspects of models frequently operate on a flawed premise. For example, Sap et al. (2022) demonstrated that although we can administer ToM tests to LLMs, the question itself is ill-posed. Humans' ToM can be gauged via question-based psychological tests because their responses are influenced by their complex inner workings. In contrast, LLMs respond by generating a list of likely words. No ToM required. Similarly, Shu et al. (2024) demonstrate that while LLMs can generate answers to psychometric questionnaires like personality tests, their answers are inconsistent and lack awareness of the premise. Even when human and model responses are similar, they stem from very different causes. In the absence of explicit modeling, it is unclear whether LLMs would develop social capabilities by themselves.

As we enter the era of LLM-dominated NLP, the next logical step is to tackle "harder" problems. Applying Moravec's paradox, the next more difficult area for NLP would be either motion (less applicable and addressed by robotics, but possible in the context of multimodality) or social awareness. This step aligns with a growing societal need. However, making progress in this area means answering difficult questions: Is it possible to gain social awareness gradually and/or systematically? Can we teach our models how humans develop social awareness? Despite the difficulty of replicating human social awareness in machines, we advocate for the development of NLP systems capable of learning and recognizing social awareness over time, as well as responding to these cues in a more human-like manner.

6. The Future of (Socially Aware) Language Technologies

NLP is not the first field to focus on the abstract over the concrete. Linguistics used to view language as separate from all other cognitive (and physical) abilities. While this abstract framing allowed for studying specific aspects in isolation and developing theories and models, it obscured the overall picture. Sociolinguistics, psycholinguistics,

and other subfields have worked hard to reintroduce the importance of "extraneous" factors into the linguistic mainstream.

Today, NLP follows a similar trajectory. Many traditional NLP tasks have become obsolete as LLMs play a more significant role in AI research. However, as the power of those models grows, we are increasingly free to think about their use in a techno-social environment (Blodgett et al. 2020; Tedeschi et al. 2023; Abercrombie et al. 2023). With language models providing the foundation for natural language generation and analysis, we can (re)focus on the social aspects of language modeling. Understanding the social aspects of language technologies requires a focus on emotional intelligence, cultural factors, values, norms, social interaction, and broader social implications. Developing socially aware NLP requires more than simply building models that recognize social factors, as Hovy and Yang (2021) have suggested; it also involves examining how these NLP systems interact with both social and physical environments, as well as their broad social implications. As long as NLP systems exist, social awareness will remain essential, because social factors, interactions, and their implications are integral to any human engagement with these technologies.

Socially aware NLP is likely to transform industries and societal functions while also shaping the broader field of AI, including audio, vision, and robotics, where social awareness can play an even more critical role. Integrating social awareness in robotics can enable the development of robots that can safely and effectively interact with humans (e.g., eldercare robots, service robots), and advancements in computer vision that enable systems to better interpret emotions, social interactions, and cultural contexts from visual data (Mittal et al. 2020; Kwon et al. 2023; Kruk, Ziems, and Yang 2023; Achlioptas et al. 2021). In contrast, developing socially aware NLP can also introduce significant risks such as misunderstanding cultures, enforcing biases, and violating privacy. The misuse of such socially aware systems may also lead to over-reliance and echo chambers, as well as misinformation by bad actors. We should proceed with a keen awareness of ethics and risks (Barrett et al. 2023).

In the future, we can look into how these models function as social agents, what social cues they read and understand, and what tasks requiring social awareness they can complete. This pivot will necessitate new tasks, metrics, and approaches fundamentally different from the goals we have pursued as a field thus far. Most importantly, it will necessitate a re-alignment of the current fractured AI landscape: We will need to collaborate across disciplines to incorporate social awareness into our models. There are numerous unexplored research areas awaiting exploration.

Humans are more than language factories. Language is only one component of our complex social interactions. We are not human because we speak—we speak because we are human. Language models, on the other hand, and at this point, are language factories capable of producing and processing words at astonishing rates but lacking the faculties that drive human language production and processing of world knowledge and social nuances. Socially aware language technologies can get us closer to AI's initial goals, advance the field, and help address many of the current issues we face.

References

Abercrombie, Gavin, Amanda Curry, Tanvi Dinkar, Verena Rieser, and Zeerak Talat. 2023. Mirages. On anthropomorphism in dialogue systems. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4776–4790. https://doi.org/10.18653/v1/2023.emnlp-main.290

Achlioptas, Panos, Maks Ovsjanikov, Kilichbek Haydarov, Mohamed Elhoseiny, and Leonidas J. Guibas. 2021. ArtEmis:

Affective language for visual art. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11569–11579. https://doi.org/10.1109/CVPR46437.2021.01140

Barrett, Clark, Brad Boyd, Elie Bursztein, Nicholas Carlini, Brad Chen, Jihye Choi, Amrita Roy Chowdhury, Mihai Christodorescu, Anupam Datta, Soheil Feizi, et al. 2023. Identifying and mitigating the security risks of generative AI. *Foundations and Trends in Privacy and Security*, 6(1):1–52. https://doi.org/10.1561/9781638283133

Bianchi, Federico, Mirac Suzgun, Giuseppe Attanasio, Paul Röttger, Dan Jurafsky, Tatsunori Hashimoto, and James Zou. 2023. Safety-tuned LLaMAs: Lessons from improving the safety of large language models that follow instructions. *arXiv preprint arXiv:2309.07875*.

Bird, Steven and Dean Yibarbuk. 2024. Centering the speech community. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 826–839.

Blair, Irene V., Jennifer E. Ma, and Alison P. Lenton. 2001. Imagining stereotypes away: The moderation of implicit stereotypes through mental imagery. *Journal of Personality and Social Psychology*, 81(5):828–841. https://doi.org/10.1037/0022-3514.81.5.828, PubMed: 11708560

Blodgett, Su Lin, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476. https://doi.org/10.18653/v1/2020.acl-main.485

Bolukbasi, Tolga, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam T. Kalai. 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. *Advances in Neural Information Processing Systems*, 29:4356–4364.

Breazeal, Cynthia. 2003. Emotion and sociable humanoid robots. *International Journal of Human-computer Studies*, 59(1–2):119–155. https://doi.org/10.1016/S1071-5819(03)00018-1

Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

Choi, Minje, Ceren Budak, Daniel M. Romero, and David Jurgens. 2021. More than meets the tie: Examining the role of interpersonal relationships in social networks. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, pages 105–116. https://doi.org/10.1609/icwsm.v15i1.18045

Choi, Minje, Jiaxin Pei, Sagar Kumar, Chang Shu, and David Jurgens. 2023. Do LLMs understand social knowledge? Evaluating the sociability of large language models with socket benchmark. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11370–11403. https://doi.org/10.18653/v1/2023.emnlp-main.699

Colman, Andrew M. 2003. Cooperation, psychological game theory, and limitations of rationality in social interaction. *Behavioral and Brain Sciences*, 26(2):139–153. https://doi.org/10.1017/S0140525X03000050, PubMed: 14621510

Cooley, Charles Horton. 1902. The looking-glass self. *The Production of Reality: Essays and Readings on Social Interaction*, 6(1902):126–128.

Dev, Sunipa, Vinodkumar Prabhakaran, David Adelani, Dirk Hovy, and Luciana Benotti, editors. 2023. *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*.

Dev, Sunipa, Emily Sheng, Jieyu Zhao, Aubrie Amstutz, Jiao Sun, Yu Hou, Mattie Sanseverino, Jiin Kim, Akihiro Nishi, Nanyun Peng, et al. 2022. On measures of biases and harms in NLP. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 246–267. https://doi.org/10.18653/v1/2022.findings-aacl.24

Dhuliawala, Shehzaad, Vilém Zouhar, Mennatallah El-Assady, and Mrinmaya Sachan. 2023. A diachronic perspective on user trust in AI under uncertainty. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5567–5580. https://doi.org/10.18653/v1/2023.emnlp-main.339

Dinan, Emily, Gavin Abercrombie, A. Bergman, Shannon Spruit, Dirk Hovy, Y-Lan Boureau, and Verena Rieser. 2022. SafetyKit: First aid for measuring safety in open-domain conversational systems. In *Proceedings of the 60th Annual Meeting of the*

Association for Computational Linguistics (Volume 1: Long Papers), pages 4113–4133. `https://doi.org/10.18653/v1/2022 .acl-long.284`

Dohmen, Thomas, Armin Falk, David Huffman, Uwe Sunde, Jürgen Schupp, and Gert G. Wagner. 2005. Individual risk attitudes: New evidence from a large, representative, experimentally-validated survey. Technical report, DIW Discussion Papers. `https://doi.org/10.2139 /ssrn.807408`

Eloundou, Tyna, Sam Manning, Pamela Mishkin, and Daniel Rock. 2023. GPTs are GPTs: An early look at the labor market impact potential of large language models. *arXiv preprint arXiv:2303.10130.* `https:// doi.org/10.1126/science.adj0998`, PubMed: 38900883

ElSherief, Mai, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. Latent hatred: A benchmark for understanding implicit hate speech. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 345–363. `https://doi.org/10.18653/v1/2021 .emnlp-main.29`

Evans, Anthony M. and Joachim I. Krueger. 2009. The psychology (and economics) of trust. *Social and Personality Psychology Compass*, 3(6):1003–1017. `https://doi .org/10.1111/j.1751-9004.2009 .00232.x`

Fathi, Alircza, Jessica K. Hodgins, and James M. Rehg. 2012. Social interactions: A first-person perspective. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1226–1233. `https:// doi.org/10.1109/CVPR.2012.6247805`

Flek, Lucie. 2020. Returning the N to NLP: Towards contextually personalized classification models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7828–7838. `https://doi.org/10.18653/v1/2020 .acl-main.700`

Gandhi, Kanishk, Jan-Philipp Fränken, Tobias Gerstenberg, and Noah Goodman. 2024. Understanding social reasoning in language models with language models. *Advances in Neural Information Processing Systems*, 36.

Gehman, Samuel, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. *arXiv preprint arXiv:2009.11462.* `https://doi.org /10.18653/v1/2020.findings-emnlp .301`

Goleman, Daniel. 2006. *Social Intelligence: The New Science of Human Relationships*. Bantam Dell Publishing Group.

Gonen, Hila and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 609–614.

Grant, Erin, Aida Nematzadeh, and Thomas L. Griffiths. 2017. How can memory-augmented neural networks pass a false-belief task? In *CogSci*. `https://api.semanticscholar.org /CorpusID:7340345`

Guo, Anhong, Ece Kamar, Jennifer Wortman Vaughan, Hanna Wallach, and Meredith Ringel Morris. 2020. Toward fairness in AI for people with disabilities SBG@a research roadmap. *ACM SIGACCESS Accessibility and Computing*, (125):1–1. `https://doi.org/10.1145/ 3386296.3386298`

Held, William, Camille Harris, Michael Best, and Diyi Yang. 2023. A material lens on coloniality in NLP. *arXiv preprint arXiv:2311.08391.*

Hershcovich, Daniel, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. 2022. Challenges and strategies in cross-cultural NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013. `https://doi.org/10.18653/v1/2022 .acl-long.482`

Hovy, Dirk, Federico Bianchi, and Tommaso Fornaciari. 2020. "You sound just like your father". Commercial machine translation systems include stylistic biases. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1686–1690. `https://doi.org/10 .18653/v1/2020.acl-main.154`

Hovy, Dirk and Shannon L. Spruit. 2016. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short*

Papers), pages 591–598. `https://doi.org` `/10.18653/v1/P16-2096`

Hovy, Dirk and Diyi Yang. 2021. The importance of modeling social factors of language: Theory and practice. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 588–602. `https://doi.org/10.18653/v1/2021` `.naacl-main.49`

Huang, Jie and Kevin Chen-Chuan Chang. 2023. Towards reasoning in large language models: A survey. In *61st Annual Meeting of the Association for Computational Linguistics, ACL 2023*, pages 1049–1065. `https://` `doi.org/10.18653/v1/2023.findings` `-acl.67`

Ivey, Jonathan, Shivani Kumar, Jiayu Liu, Hua Shen, Sushrita Rakshit, Rohan Raju, Haotian Zhang, Aparna Ananthasubramaniam, Junghwan Kim, Bowen Yi, et al. 2024. Real or robotic? Assessing whether LLMs accurately simulate qualities of human responses in dialogue. *arXiv preprint arXiv:2409.08330*.

Iyyer, Mohit, Anupam Guha, Snigdha Chaturvedi, Jordan Boyd-Graber, and Hal Daumé III. 2016. Feuding families and former friends: Unsupervised learning for dynamic fictional relationships. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1534–1544. `https://` `doi.org/10.18653/v1/N16-1180`

Jakesch, Maurice, Advait Bhat, Daniel Buschek, Lior Zalmanson, and Mor Naaman. 2023. Co-writing with opinionated language models affects users' views. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–15. `https://doi.org` `/10.1145/3544548.3581196`

Jin, Zhijing, Geeticka Chauhan, Brian Tse, Mrinmaya Sachan, and Rada Mihalcea. 2021. How good is NLP? A sober look at NLP tasks through the lens of social impact. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3099–3113. `https://doi.org/10` `.18653/v1/2021.findings-acl.273`

Joshi, Pratik, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*,

pages 6282–6293. `https://doi.org/10` `.18653/v1/2020.acl-main.560`

Kantharuban, Anjali, Jeremiah Milbauer, Emma Strubell, and Graham Neubig. 2024. Stereotype or personalization? User identity biases chatbot recommendations. *arXiv preprint arXiv:2410.05613*.

Kruk, Julia, Caleb Ziems, and Diyi Yang. 2023. Impressions: Visual semiotics and aesthetic impact understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12273–12291. `https://doi.org/10.18653/v1/2023` `.emnlp-main.755`

Kumar, Deepak, Yousef Anees AbuHashem, and Zakir Durumeric. 2024. Watch your language: Investigating content moderation with large language models. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, pages 865–878. `https://` `doi.org/10.1609/icwsm.v18i1` `.31358`

Kwon, Minae, Hengyuan Hu, Vivek Myers, Siddharth Karamcheti, Anca Dragan, and Dorsa Sadigh. 2023. Toward grounded social reasoning. *arXiv preprint arXiv:2306.08651*. `https://doi.org/10` `.1109/ICRA57147.2024.10611218`

Lapinski, Maria Knight and Rajiv N. Rimal. 2005. An explication of social norms. *Communication Theory*, 15(2):127–147. `https://doi.org/10.1111/j.1468` `-2885.2005.tb00329.x`

Lauscher, Anne, Federico Bianchi, Samuel Bowman, and Dirk Hovy. 2022. SocioProbe: What, when, and where language models learn about sociodemographics. *arXiv preprint arXiv:2211.04281*. `https://doi.org` `/10.18653/v1/2022.emnlp-main.539`

Le, Matthew, Y-Lan Boureau, and Maximilian Nickel. 2019. Revisiting the evaluation of theory of mind through question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5872–5877. `https://doi.org/10` `.18653/v1/D19-1598`

Litschko, Robert, Max Müller-Eberstein, Rob Van Der Goot, Leon Weber, and Barbara Plank. 2023. Establishing trustworthiness: Rethinking tasks and model evaluation. *arXiv preprint arXiv:2310.05442*. `https://doi.org/10.18653/v1/2023` `.emnlp-main.14`

Liu, Yihe, Anushk Mittal, Diyi Yang, and Amy Bruckman. 2022. Will AI console me when I lose my pet? Understanding perceptions of AI mediated email writing. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–13. `https://doi.org/10.1145/3491102.3517731`

Ma, Bolei, Xinpeng Wang, Tiancheng Hu, Anna-Carolina Haensch, Michael A. Hedderich, Barbara Plank, and Frauke Kreuter. 2024. The potential and challenges of evaluating attitudes, opinions, and values in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8783–8805. `https://doi.org/10.18653/v1/2024.findings-emnlp.513`

McCarthy, John, Marvin L. Minsky, Nathaniel Rochester, and Claude E. Shannon. 2006. A proposal for the Dartmouth summer research project on artificial intelligence, August 31, 1955. *AI Magazine*, 27(4):12–12.

Mittal, Trisha, Pooja Guhan, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, and Dinesh Manocha. 2020. EmotiCon: Context-aware multimodal emotion recognition using Frege's Principle. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14234–14243. `https://doi.org/10.1109/CVPR42600.2020.01424`

Moravec, Hans. 1988. *Mind children: The future of robot and human intelligence*. Harvard University Press.

Nguyen, Dong, A. Seza Doğruöz, Carolyn P. Rosé, and Franciska de Jong. 2016. Computational sociolinguistics: A Survey. *Computational Linguistics*, 42(3):537–593. `https://doi.org/10.1162/COLI_a_00258`

Nguyen, Dong, Laura Rosseel, and Jack Grieve. 2021. On learning and representing social meaning in NLP: A sociolinguistic perspective. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 603–612. `https://doi.org/10.18653/v1/2021.naacl-main.50`

Park, Joon Sung, Carolyn Q. Zou, Aaron Shaw, Benjamin Mako Hill, Carrie Jun Cai, Meredith Ringel Morris, Robb Willer, Percy Liang, and Michael S. Bernstein. 2024. Generative agent simulations of 1,000 people. *arXiv preprint arXiv:2411.10109v1*.

Pawar, Siddhesh, Junyeong Park, Jiho Jin, Arnav Arora, Junho Myung, Srishti Yadav, Faiz Ghifari Haznitrama, Inhwa Song, Alice Oh, and Isabelle Augenstein. 2024. Survey of cultural awareness in language models: Text and beyond. *arXiv preprint arXiv:2411.00860*.

Pentland, Alex. 2005. Socially aware computation and communication. In *Proceedings of the 7th International Conference on Multimodal Interfaces*, pages 199–199. `https://doi.org/10.1145/1088463.1088466`

Pinker, Steven. 2003. *The Language Instinct: How the Mind Creates Language*. Penguin UK.

Prabhakaran, Vinodkumar, Ajita John, and Dorée D. Seligmann. 2013. Who had the upper hand? Ranking participants of interactions based on their relative power. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 365–373.

Premack, David and Guy Woodruff. 1978. Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, 1(4):515–526. `https://doi.org/10.1017/S0140525X00076512`

Ranathunga, Surangika and Nisansa de Silva. 2022. Some languages are more equal than others: Probing deeper into the linguistic disparity in the NLP world. *arXiv preprint arXiv:2210.08523*. `https://doi.org/10.18653/v1/2022.aacl-main.62`

Rathje, Steve, Dan-Mircea Mirea, Ilia Sucholutsky, Raja Marjieh, Claire E. Robertson, and Jay J. Van Bavel. 2024. GPT is an effective tool for multilingual psychological text analysis. *Proceedings of the National Academy of Sciences*, 121(34):e2308950121. `https://doi.org/10.1073/pnas.2308950121`, PubMed: 39133853

Rhodes, Marjorie and Andrew Baron. 2019. The development of social categorization. *Annual Review of Developmental Psychology*, 1(1):359–386. `https://doi.org/10.1146/annurev-devpsych-121318-084824`, PubMed: 33103119

Ryan, Michael J., William Held, and Diyi Yang. 2024. Unintended impacts of LLM alignment on global representation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16121–16140. `https://doi.org/10.18653/v1/2024.acl-long.853`

Sap, Maarten, Ronan Le Bras, Daniel Fried, and Yejin Choi. 2022. Neural theory-of-mind? On the limits of social intelligence in large LMs. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3762–3780. `https://doi.org/10.18653/v1/2022.emnlp-main.248`

Shah, Deven Santosh, H. Andrew Schwartz, and Dirk Hovy. 2020. Predictive biases in natural language processing models: A conceptual framework and overview. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5248–5264. `https://doi.org/10.18653/v1/2020.acl-main.468`

Sharma, Ashish, Kevin Rushton, Inna Wanyin Lin, David Wadden, Khendra G. Lucas, Adam S. Miner, Theresa Nguyen, and Tim Althoff. 2023. Cognitive reframing of negative thoughts through human-language model interaction. *arXiv preprint arXiv:2305.02466*. `https://doi.org/10.18653/v1/2023.acl-long.555`

Sher, Itai, Melissa Koenig, and Aldo Rustichini. 2014. Children's strategic theory of mind. *Proceedings of the National Academy of Sciences*, 111(37):13307–13312. `https://doi.org/10.1073/pnas.1403283111`, PubMed: 25197065

Shu, Bangzhao, Lechen Zhang, Minje Choi, Lavinia Dunagan, Dallas Card, and David Jurgens. 2024. You don't need a personality test to know these models are unreliable: Assessing the reliability of large language models on psychometric instruments. In *Proceedings of the 2024 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. `https://doi.org/10.18653/v1/2024.naacl-long.295`

Snyder, Mark and William Ickes. 1985. Personality and social behavior. *Handbook of Social Psychology*, 2(3):883–947.

Song, Yueqi, Catherine Cui, Simran Khanuja, Pengfei Liu, Fahim Faisal, Alissa Ostapenko, Genta Indra Winata, Alham Fikri Aji, Samuel Cahyawijaya, Yulia Tsvetkov, et al. 2023. Globalbench: A benchmark for global progress in natural language processing. *arXiv preprint arXiv:2305.14716*. `https://doi.org/10.18653/v1/2023.emnlp-main.875`

Soni, Nikita, Lucie Flek, Ashish Sharma, Diyi Yang, Sara Hooker, and H. Andrew Schwartz, editors. 2024. *Proceedings of the 1st Human-Centered Large Language Modeling Workshop*.

Sorensen, Taylor, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, et al. 2024. A roadmap to pluralistic alignment. *arXiv preprint arXiv:2402.05070*.

Spangher, Alexander, Nanyun Peng, Sebastian Gehrmann, and Mark Dredze. 2024. Do LLMs plan like human writers? Comparing journalist coverage of press releases with LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21814–21828. `https://doi.org/10.18653/v1/2024.emnlp-main.1216`

Suh, Jina, Tim Althoff, and John Torous. 2024. Special report: Are you ready for generative AI in psychiatric practice? *Psychiatric News*, 59(11). `https://doi.org/10.1176/appi.pn.2024.11.11.10`

Tedeschi, Simone, Johan Bos, Thierry Declerck, Jan Hajič, Daniel Hershcovich, Eduard Hovy, Alexander Koller, Simon Krek, Steven Schockaert, Rico Sennrich, Ekaterina Shutova, and Roberto Navigli. 2023. What's the meaning of superhuman performance in today's NLU? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12471–12491. `https://doi.org/10.18653/v1/2023.acl-long.697`

Thompson, Ross A. 2007. The development of the person: Social understanding, relationships, conscience, self. *Handbook of Child Psychology*, 3. `https://doi.org/10.1002/9780470147658.chpsy0302`

Tian, Yufei, Tenghao Huang, Miri Liu, Derek Jiang, Alexander Spangher, Muhao Chen, Jonathan May, and Nanyun Peng. 2024. Are large language models capable of generating human-level narratives? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17659–17681. `https://doi.org/10.18653/v1/2024.emnlp-main.978`

Tomasello, Michael. 2014. *A natural history of human thinking*. Harvard University Press. `https://doi.org/10.4159/9780674726369`

Turing, Alan M. 1950. I. Computing machinery and intelligence. *Mind*, LIX(236):433–460. `https://doi.org/10.1093/mind/LIX.236.433`

Vanmassenhove, Eva, Christian Hardmeier, and Andy Way. 2018. Getting gender right in neural machine translation. In *Proceedings of the 2018 Conference on*

Empirical Methods in Natural Language Processing, pages 3003–3008. `https://doi.org/10.18653/v1/D18-1334`

Vida, Karina, Fabian Damken, and Anne Lauscher. 2024. Decoding multilingual moral preferences: Unveiling LLM's biases through the moral machine experiment. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pages 1490–1501. `https://doi.org/10.1609/aies.v7i1.31741`

Wang, Rose E., Qingyang Zhang, Carly Robinson, Susanna Loeb, and Dorottya Demszky. 2024. Bridging the novice-expert gap via models of decision-making: A case study on remediating math mistakes. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 2174–2199. `https://doi.org/10.18653/v1/2024.naacl-long.120`

Wilson, Kyra and Aylin Caliskan. 2024. Gender, race, and intersectional bias in resume screening via language model retrieval. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pages 1578–1590. `https://doi.org/10.1609/aies.v7i1.31748`

Wright, Dustin, Arnav Arora, Nadav Borenstein, Srishti Yadav, Serge Belongie, and Isabelle Augenstein. 2024. LLM tropes: Revealing fine-grained values and opinions in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 17085–17112. `https://doi.org/10.18653/v1/2024.findings-emnlp.995`

Yin, Kayo, Amit Moryossef, Julie Hochgesang, Yoav Goldberg, and Malihe Alikhani. 2021. Including signed languages in natural language processing. *arXiv preprint arXiv:2105.05222*. `https://doi.org/10.18653/v1/2021.acl-long.570`

Zhang, Jintian, Xin Xu, Ningyu Zhang, Ruibo Liu, Bryan Hooi, and Shumin Deng. 2023. Exploring collaboration mechanisms for LLM agents: A social psychology view. *arXiv preprint arXiv:2310.02124*.

Ziems, Caleb, Omar Shaikh, Zhehao Zhang, William Held, Jiaao Chen, and Diyi Yang. 2023. Can large language models transform computational social science? *Computational Linguistics*, 50(1):237–291. `https://doi.org/10.1162/coli_a_00502`