Socially Aware Language Technologies: Perspectives and Practices

Anonymous ACL submission

Abstract

Language technologies have made enormous progress, especially with the introduction of large language models (LLMs). These advances can, however, exacerbate a variety of issues that models have traditionally struggled with, such as bias, evaluation, and risks. In this perspective paper, we argue that many of these issues share a common core: a lack of awareness of the factors, context, and implications of the social environment in which NLP operates, which we call **social awareness**. While NLP is getting better at solving the linguistic aspects, relatively limited progress has been made in adding the social awareness required for language applications to work in all situations for all users. Integrating social awareness into NLP models will make applications more natural, helpful, and safe, and will open up new possibilities. Thus we argue that substantial challenges remain for NLP to develop social awareness and that we are just at the beginning of a new era for the field.

1 Introduction

Natural language processing (NLP) has made significant strides in recent years, thanks in part to the introduction of large pretrained language models (LLMs) based on Transformers (Vaswani et al., 2017; Brown et al., 2020). As a result, performance on various NLP tasks, such as machine translation, sentiment analysis, or conversational agents, has significantly improved. These models now perform these tasks seemingly as well as, if not better than, humans (Tedeschi et al., 2023). On the other hand, a growing number of issues and shortcomings with these models has been reported. Some of these issues include bias (Bolukbasi et al., 2016), toxicity (Gehman et al., 2020), trust (Litschko et al., 2023), and concerns about fairness (Hovy and Spruit, 2016; Blodgett et al., 2020; Shah et al., 2020; ElSherief et al., 2021). Word embeddings, which represent words in a mathematical space,

can, for example, inadvertently capture and reinforce biases in training data, perpetuating stereotypes and inequalities (Bolukbasi et al., 2016; Gonen and Goldberg, 2019). Machine translation systems have been shown to generate translations with unintended biases or inaccuracies (Vanmassenhove et al., 2018; Hovy et al., 2020), potentially exacerbating cultural and societal misunderstandings (Bird and Yibarbuk, 2024). Furthermore, NLP applications are still insufficient for tasks that require social awareness, especially in high-stakes areas like health care. In its current form, NLP only serves a subset of people and situations that use language technology (Held et al., 2023).

Many of these issues facing modern NLP share a common core. Namely, they result from failing to consider language (technologies) in the context of communities, cultural and ideological differences, and social contexts. All these social awareness aspects are relevant not just for English, but also the 7,000 languages out there (Joshi et al., 2020), adding complexity to the problem. These issues fall under social awareness:

> **Social Awareness** refers to the ability to be aware of social social factors, contexts, and social dynamics, as well as the implications of language use on the broader social environment.

Social awareness is undervalued in current NLP. Very traditional NLP models often focus on syntax, grammar, and lexicon, but have not made much progress in capturing cultural context and social interactions. The inherent difficulty of operationalizing and integrating these complexities into today's LLMs is a significant reason. We argue that we need to address this issue to take NLP to the next level. We must broaden the scope of NLP technologies to a wider range of people and situations, and advance and promote inclusivity and accessibility across different languages and cultures (Hovy and

Yang, 2021; Hershcovich et al., 2022). This is a message that needs to be **repeated** in the NLP community because we are nearing a point where such social awareness can be integrated into systems.

Social awareness is not restricted to NLP; it should be an integral and foundational component across all modalities of AI. While our view applies most readily to NLP, it is also relevant for vision (Fathi et al., 2012) and robotics (Breazeal, 2003), for example. Social awareness governs the dynamics of human-human and human-AI interactions and has an impact on knowledge acquisition and use. Language, as a means of communication, serves as a tool for individuals to achieve a variety of goals, even though it is generated and consumed by people from various backgrounds. NLP's potential insights and applications will inevitably be limited if it does not consider the interaction of individuals, the context in which language is uttered, and the specific goals it should achieve. Knowledge of such goals or capabilities in turn enables users to gain more trust in NLP system— social awareness is also an important factor for more trustworthy NLP in the future (Litschko et al., 2023). This is because language is more than just words and grammar; human society and culture is inextricably linked to it. By modeling the social factors that influence language, our AI systems can better understand and connect with people, and expand their scope and depth. Concretely, in this position paper, we introduce three key aspects socially aware NLP needs to account for to work, namely **social factors** (Section 2.1), **social interaction** (Section 2.2), and **social implication** (Section 2.3).

2   What Is Socially Aware NLP?

Pentland (2005) defines *socially aware computation* as systems that can understand social signaling and social context. The author argues that focusing on such dimensions can improve collective decision-making and keep users informed. From a psychology perspective, Daniel Goleman defines a related term—*emotional intelligence*,—-which he breaks down into four subsets: self-awareness, self-management, social awareness, and relationship management (Hernez-Broome, 2012). For emotional intelligence, social awareness requires the ability to accurately understand other people's emotions and empathize with them, which relates to having a Theory of Mind of others (Tomasello, 2014; Premack and Woodruff, 1978).

The concept of **social awareness** in natural language understanding refers to shifting focus away from classic tasks and benchmarks and instead encouraging tasks, models, and evaluations to consider social factors (Hovy and Yang, 2021), social context, and social dynamics communicated through language. The "awareness" pertains to language technologies themselves being designed to recognize and exhibit such social aspects and process these socially driven meanings and implications behind language as humans do. In addition to the increased awareness of language technologies, researchers or practitioners should also be aware of these social aspects to design language technologies that are socially aware.

> **Socially aware language technologies** refer to the study and development of language technologies from a social perspective, to provide NLP systems with the ability to understand the social context, perspectives, and emotions expressed in language by humans.

In other words, a socially aware system should demonstrate emotional intelligence, social intelligence, cultural competence, or perspective-taking abilities. Consequently, research in socially aware language technologies will focus on developing new algorithms, models, evaluation metrics, and approaches that allow NLP systems to be more socially aware and to better recognize and respond to social cues, cultural nuances, and other human communication-related factors. This research also entails the design and implementation of socially-aware NLP systems for use in real-world applications such that their use, implications, and impact are all understood during their development.

2.1   Social Factors

Hovy and Yang (2021) lay out seven *social factors* that NLP systems need to incorporate: *speaker* characteristics, *receiver* characteristics, *social relations*, *context*, *social norms*, *culture and ideology*, and *communicative goals*. These aspects provide an easy-to-use conceptual taxonomy on different social factors in language. It is also essential to consider the incorporation and understanding of social factors in various domains, including psychology, sociology, and more.

> **Social factors** refers to a wide range of social aspects that shape the way we understand language use, including but not
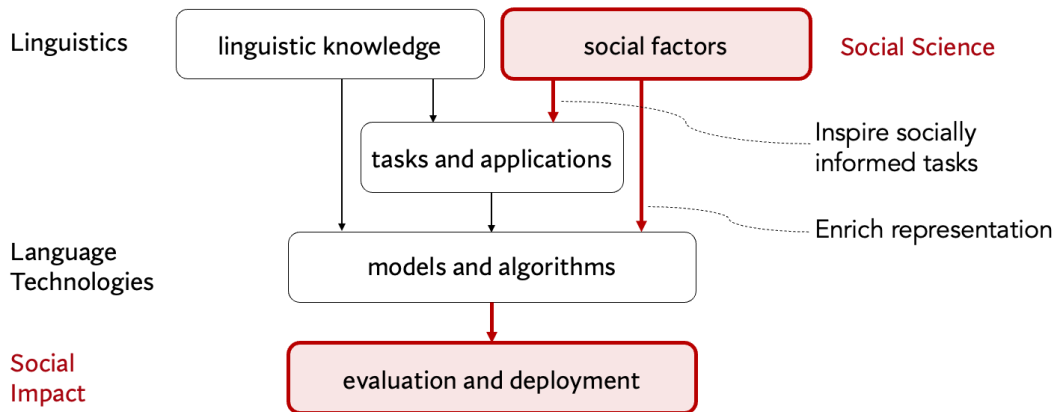
Figure 1: How social factors relates to language technologies, linguistics and social sciences

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 which communicative goals.

Figure 1 (without the highlighted red boxes) shows how current NLP works. We typically have some tasks inspired by linguistic knowledge and then build and evaluate models for the task, such as Natural Language Inference (Bowman et al., 2015). To develop socially aware language technologies, we have to incorporate social factors from social science into this pipeline. Social factors, in particular, can inspire socially informed tasks through additional objective functions and tasks. Operationalizing social phenomena will augment the current pool of tasks to better reflect users' needs and likely lead to increased user trust. Furthermore, social knowledge can supplement existing representations in models (Nguyen et al., 2021). Social signals may provide alternative supervision for representation learning and next-word prediction. Current models have internal representations of social factors but do not appear to actively draw on them (Lauscher et al., 2022). With social awareness integrated into the pipeline, the outcome can produce social impact, not only about the typical task evaluation metric but also the impact on people.

2.2 Social Interaction

To develop socially aware language technologies, we must go beyond social factors (Hovy and Yang, 2021) and consider a wide range of **social interactions** and their **social implications**, such as the relationship, organizational, and cultural norms that govern interpersonal communication. While social factors intrinsically encompass language, culture, and behavior, the awareness of social interactions and implications digs into the *dynamic context* in which NLP systems operate and their impact.

In social science, many theories, such as those surrounding social influence and norms, define essential dimensions of social interaction, shedding light on the intricate processes underlying human behavior and interaction. The desire to communicate within this environment helps drive language development, and our environment also profoundly influences the speed and efficiency of our language acquisition (Plunkett, 1997). This perspective posits that language is not an isolated 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 regarded as groups' shared standards of acceptable behavior. Some social norms become laws and rules, while others remain informal but equally influential. Integrating social norms into language introduces a layer of complexity beyond mere vocabulary and grammar. The work of Lapinski and Rimal (2005) highlights the nuanced interplay between social norms and language, demonstrating that linguistic expressions often serve as vehicles for the expression and reinforcement of these norms. All of these provide a rich and multifaceted foundation for our explorations of the complicated space of social interaction.

Social interaction refers to the interaction dynamics, including social exchange between individuals, other people in the

3

context who are involved and other activities surrounding the context.

Similar to *how we see ourselves comes from our perception of how others see us*, language use is influenced by others' views of the language, especially as people increasingly interact with LLMs. Socio-technical NLP systems exist within an ecosystem of social interactions, where users, developers, and stakeholders come together to create, deploy, and use these technologies. Many factors influence these social interactions, including power dynamics (Prabhakaran et al., 2013), trust (Litschko et al., 2023), and user expectations (Dhuliawala et al., 2023). The design of NLP systems must consider how these social interactions shape user experiences (Jakesch et al., 2023; Liu et al., 2022b) and how they impact the adoption and effectiveness of the technology.

2.3 Social Implication

Social implication encompasses a complex interplay of positive and negative effects of NLP systems on society. Understanding the effects and consequences of language technologies on society is crucial to the responsible development and sustainable use of NLP, which includes but not limited to assessing biases and stereotypes (Dev et al., 2022); considering how systems affect global populations, not just those in the North (Song et al., 2023; Ranathunga and de Silva, 2022); investigating the dissemination of misinformation and dual use of NLP systems; examining concerns around job displacement (Eloundou et al., 2023) and understanding how model uses are associated with creativity, trust, and productivity (Chakrabarty et al., 2024; Liu et al., 2022b).

> **Social implication** refers to broadly understanding an NLP system's implications and impact on society.

Aside from understanding such broad implications of NLP systems, social implications can also inspire model design, such as designing models to consider the implications of their outputs. For example, work in prompt safety can be seen as an initial step at imbuing models with a sense of what responses have harmful social implications and teaching them to decline to answer (e.g., Bianchi et al., 2023). As a society, we need to understand these social implications if we are to use NLP to its full potential and mitigate harms.
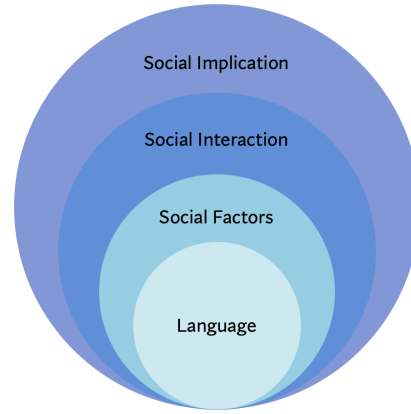
Figure 2: Conceptual structure of socially aware NLP: social factors, social interaction, and social implication. Note that this is **not an exclusive partition**, but one way for us to understand the scope of different tasks.

Figure 2 shows that in development, social awareness is first present in the initial design of the task and algorithms (social factors). It plays a role in the middle ground of interactions and activities around it (social interactions) and the outer layer of impact (social implication). By putting a strong emphasis on social factors, social interaction, and their social implications, we hope socially aware language technologies can foster development that improves communication and aligns with ethical and societal values, promoting a more harmonious and equitable sociotechnical ecosystem.

2.4 Emerging Work in Socially Aware NLP

An increasing amount of research has been conducted around socially aware NLP.

From a **social factor** perspective, research has attempted to understand speaker and receiver characteristics (Flek, 2020; Hovy et al., 2020), such as personalized text generation (Wu et al., 2021) or demographic aspects (Hovy, 2015); model social relations to improve language understanding (Yang et al., 2019b; Iyyer et al., 2016); analyze culture for more contextualized algorithms (Huang and Yang, 2023; Kozlowski et al., 2018; Garimella et al., 2019); align NLP systems with values and preferences (Durmus et al., 2023; Liu et al., 2022a; Bai et al., 2022); and investigate the broader goals behind language interaction(Stab and Gurevych, 2014; Yang et al., 2019a).

From a **social interaction** perspective, recent work around LLMs has examined how humans interact with LLMs to accomplish tasks (Yuan et al., 2022; Shaikh et al., 2023a), and to simulate human interactions (Liu et al., 2023; Park et al., 2023; Aher

4

et al., 2023), building on earlier studies that used how humans interact to develop social or linguistic theories such as common ground (Shaikh et al., 2023b; Li et al., 2023; Paranjape and Manning, 2021; Pilán et al., 2023).

From a **social implication** perspective, a growing amount of research has started to look at the implications of language technologies, such as how they affect the labor market (Eloundou et al., 2023; Chen et al., 2023), how they assist in the preservation of languages and dialects (Sap et al., 2019; Ziems et al., 2022b) (see Meta's efforts to develop high-quality machine translation capabilities for most of the world's languages)[1], and how they can be used to support education (Kasneci et al., 2023), crisis management (Goecks and Waytowich, 2023) and accessibility (Gadiraju et al., 2023).

3   What Is Not Socially Aware NLP

As with many other things, we can best identify social awareness 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 with inappropriate or hurtful responses (e.g., telling a suicidal user to kill themselves (Dinan et al., 2022)). Here, socially aware NLP differs from personalization as it aims to incorporate a broader context of language use, such as broader social and cultural groups, while personalization focuses more on the individual for customized user experience.

Socially aware NLP and "*NLP in a social context*" are related but *not* the same concept. The use of NLP techniques to analyze and understand language use *in social settings* such as online communities, political discourse, and public opinion is referred to as NLP in a social context or sometimes through the text lens of Computational Social Science (CSS), which often develops NLP models in service of uncovering patterns and trends in text in order to answer questions in the social sciences. In contrast, *socially aware NLP* refers to NLP models' ability to recognize and respond to social factors, context, and cultural nuances in human language *in all situations*. It entails developing NLP models that recognize and respond to social cues such as tone of voice, sarcasm, humor, and cultural references to provide more accurate and effective mod-

els for understanding and communicating.

Compared to NLP in social context, CSS has a broader scope in using computational approaches (e.g., via networks, text, simulation or experiments) to analyze and model social phenomena, as well as understanding human behaviors and social sciences (Lazer et al., 2009). The focus of CSS is primarily on modeling and understanding social phenomena through computation, whereas socially aware NLP aims to refine language technologies that take social and cultural considerations into account. The other related concept is theory of mind (ToM; Grant et al., 2017; Le et al., 2019; Sap et al., 2022), which refers to the ability for models to reason about the mental state (e.g., intents, emotions, or beliefs) of others. While ToM and socially aware NLP share an overarching goal of improving models' ability to interact with humans in a more understandable and socially appropriate manner, ToM differs from socially aware NLP by attempting to mimic human-like understandings of others' mental states by attributing mental states and understanding intentions.

Human-Centered NLP and socially aware NLP both emphasize how to make NLP aware of human factors and aligned with real-world needs and contexts, as well as concerned with ethical considerations such as fairness, privacy and equality. In contrast, human-centered NLP focuses more on user-centered design to create systems tailored to user needs and is often based on iterative design, usability testing, and human in loop approaches towards improved human-system interaction, while socially aware NLP focuses on making NLP more aware of social context in which language is used and social dynamics of communication such as bias, cultural norms, and societal impact.

Similar comparisons also apply to the difference between *human-like* and *social awareness*. Human-like aims to create the type of user experience that mimics human-to-human interactions by making human-AI interactions natural and familiar to humans, while social awareness *further* encourages such interactions to be appropriate and considerate of the social environment.

4   How to Build Socially Aware NLP

We have outlined that socially aware language technologies need to understand 1) social factors, 2) social interaction, and 3) social implication to function well. However, it is not trivial to incorporate these different aspects into the development pro-

---

[1] https://ai.meta.com/research/no-language-left-behind/

cess of socially aware NLP.

4.1 Considerations for Socially Aware NLP

Building socially aware NLP requires a combination of technical and ethical considerations: (**C1**) *Access to diverse communities*: Socially aware language technologies need access to large and diverse datasets that reflect the linguistic and cultural diversity of the target user groups, which ensures that the models recognize and respond to social cue variations specific to different communities (Yin et al., 2021; Sharma et al., 2023; Wang et al., 2024). Consider standard languages or mainstream user groups, versus low-resource languages and dialects, and vulnerable populations such as older adults or people with cognitive impairments. (**C2**) *Incorporation of context and interaction dynamics*: Developing language technologies that are socially aware must incorporate context as well as interaction dynamics, including tone of voice, relevant contexts, and domain users involved in the interaction, rather than as static, standalone multiple-choice questions proxies for language understanding. (**C3**) *Ethical and social considerations*: Socially aware language technologies must be designed with ethical and social considerations, such as fairness, transparency, and privacy, to avoid perpetuating stereotypes or biases (Ma et al., 2023), and to respect user privacy. (**C4**) *Iterative design and continuous learning*: Socially aware NLP models must be continually monitored and improved to ensure they are effective and up-to-date with changing norms. This process may involve iteratively incorporating feedback from domain users and updating the models as needed to improve their accuracy and effectiveness.

4.2 Process of Building Socially Aware NLP

Let's take it a step further and explore what it means to *build socially aware language technologies*. One can start with models that identify social factors such as the speaker, the receiver, and other social contexts within a given interaction. Many existing studies have take this path, such as personalization (Wu et al., 2021) and the inference of social relations (Iyyer et al., 2016). Big data or deep learning models *alone* might not be sufficient to enable socially aware NLP, as social awareness also requires careful consideration of interaction dynamics, ethical and social issues, as well as the potential impact of NLP models on society. Socially aware NLP should be able to handle a broader range of inputs than current models, and distinguish and interpret complex social nuances effectively. If these models are successful in capturing various social implications, one could intuitively expect them to enhance downstream applications.

Socially aware language technologies may be required to not only process but also articulate the reasoning behind their predictions in a reliable manner (Sorensen et al., 2024). This requirement involves a layer of self-explanation, which explains the models' internal workings and their assumptions about social factors as they make them (Weber et al., 2023; Rajani et al., 2019). Even more than traditional models, socially aware models require transparency to ensure fair and justified outputs. Current LLMs lack such properties, and their reasoning is often non-reliable (Turpin et al., 2023; Mondorf and Plank, 2024). Thus, it is generally best to design a system to be socially aware from the beginning (e.g. Kotnis et al., 2022) instead of adding it as a patch later on. However, it is still possible to update an existing system so that it is more socially aware if care is taken to identify improvement areas, incorporate social data and context, and test and refine the system based on user feedback.

4.3 Key Directions for Socially Aware NLP

To advance socially aware NLP, we need to measure its progress. NLP has established a rich culture of tasks. We introduce multiple example broad directions around socially aware NLP to foster new approaches (see Table 1): (1) *Formulating tasks that operationalize social awareness*, which could benefit from theories and insights from social science and other related fields as discussed in Section 2. Many tasks in NLP have started to model social awareness, such as formulating the task of identifying hate speech (ElSherief et al., 2021; Breitfeller et al., 2019) or recognizing social relations (Iyyer et al., 2016; Choi et al., 2021), etc. These tasks provide representations that can be integrated to improve socially aware language understanding. (2) *Developing computational methods that detect social awareness* The goal is to develop computational models that can detect social awareness signals, as there is often an inadequate amount of supervised data for social factors, and generic NLP methods might not work well. (3) *Building systems that exhibit social awareness* Socially aware NLP aims to produce social impact by integrating social awareness into the systems' development process to consider diverse social factors. (4) *Evaluating social awareness* To evaluate interventions in the wild, we must consider in addition to quantitative

Task Description	Social Factors	Social Interaction	Social Implication	Considerations
(1) Formulate tasks that operationalize social awareness	✓			C1, C3
(2) Develop computational methods that detect social awareness	✓	✓		C2, C4
(3) Build systems that exhibit social awareness.	✓	✓	✓	C2, C3, C4
(4) Evaluate social awareness with real-world application	✓	✓	✓	C1, C2, C3
(5) Build socially aware NLP for real-world uses	✓	✓	✓	C1, C2, C3, C4
(6) Understand how socially aware NLP affects people and society	✓	✓	✓	C1, C3, C4

Table 1: Summary of representative tasks (§4.3) and their connections with different aspects of socially awareness (§2) including social factors, social interaction and social implication, as well as how these tasks should prioritize considerations that we have discussed in §4.1 for building socially aware NLP.

measures on static benchmarks or qualitative metrics like interviews. While such benchmarks show that many current systems lack simpler aspects of social awareness (Choi et al., 2023), relying solely on metrics omits the complexity and nuance of social awareness. Evaluation in the wild can help develop socially aware NLP that facilitates more human-AI interaction evaluation paradigms, and lead to new evaluation principles and protocols. (5) *Building socially aware language technologies for real-world uses* Socially aware language technologies are in a strong position to serve the diverse needs of users and contribute positively to society. One can leverage socially aware language technologies for positive impact (Jin et al., 2021) to create inclusive technologies by providing access to information and services for people with disabilities (Guo et al., 2020) or those who speak minority languages (Ziems et al., 2022a), as well as building technologies for crisis and emergency responses (Alharbi and Lee, 2022; Imran et al., 2016). (6) *Understanding how socially aware language technologies affect people and society* This essential area includes but is not limited to how such technologies affect how people communicate and interact with each other (Liu et al., 2022b); how such systems reinforce stereotypes or biases (Dev et al., 2022) and affect public trust, education, and the labor market (Eloundou et al., 2023); and how these technologies inform policy and regulation.

While not exhaustive, the above list sheds light on a few key directions for socially aware language technologies research. Some of these task categories, such as (2) and (6), have seen an increase in research lately, while others, such as (4) and (5), are still in early stages. Because more and more work is being done or needs to be done on social and language technologies, there is a crucial need for a subfield of "*socially aware language technologies*". Within this new sub-field, we need to ensure that language processing advances are both technically sophisticated and socially aware. A unified subfield focused on this goal would enable researchers to systematically address the challenges of embedding social intelligence into language models and facilitate more precise communication among scientists, policymakers, and the public. Therefore, the formal recognition of "*socially aware language technologies*" is a strategic step towards a future in which language technology responsibly intersects with human society.

5 Historical View of Socially Aware NLP

In the early days of AI, social awareness was assumed in the definitions, if not always explicitly stated (Turing, 1950; McCarthy et al., 2006). Early AI's goal was to produce *human-like* behavior, which would inevitably include a degree of social awareness and, as a result, a tighter coupling of different aspects and disciplines. AI was initially conceived in a much more holistic manner than the fragmented space suggests today. 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 difficulties matching human performance even on simple tasks (while outperforming humans on tasks that require patience or logic). Over time, NLP and other subfields of AI focused on more easily solvable tasks, which usually meant purely information-based or logical tasks that did not require any social awareness. As a consequence, NLP has spent long stretches focusing on information-heavy linguistic tasks. Only recently have social, cultural, and demographic aspects of language seen a resurgence in research (Hovy and Yang, 2021; Dev et al., 2023). However, social awareness is difficult to define and implement.

The strong performance of LLMs on a variety of language understanding tasks could initially sug-

gest that these models also have social awareness. However, many of these tasks focus on language-only problems that do not require social awareness (i.e., that could be solved without recurrence to non-textual knowledge). Further, those tasks that purported to show social, psychological, or emotional aspects of models often operate under a flawed premise. For example, Sap et al. (2022) showed that while we can administer ToM tests to LLMs, that premise is flawed. A human subject's ToM can be determined using a variety of question-based psychological tests because they answer as a result of their complex inner workings. LLMs, by contrast, respond by generating a list of likely words. Similarly, Shu et al. (2024) show that while LLMs can generate answers to psychometric questionnaires like personality tests, their answers are inconsistent and have little awareness of the premise. Thus, while the responses of humans and models are similar, they arise from very different causes. The development of these capabilities in the absence of explicitly programming will likely require new advancements that integrate social awareness in a responsible way—and corresponding new forms of evaluation that rigorously measure social awareness.

As we have reached the era in NLP dominated by LLMs, the next logical step is to tackle "harder" problems. Applying Moravec's paradox, the next harder area for NLP would involve either motion (less applicable) or emotional intelligence and social awareness. This step coincides with a growing societal need. Yet, making progress in this area requires us to begin answering hard questions: Is social awareness gained gradually and/or systematically? Can we teach our machines how humans learn social awareness? In spite of the difficulty of replicating human social awareness in machines, we call for developing NLP systems that can learn and recognize social awareness over time, and to respond to these cues in a more human-like manner.

6   The Future of (Socially Aware) NLP

As LLMs take a more central role in AI research more broadly, many traditional NLP tasks have become obsolete. However, as the information processing 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). We are not human because we speak; we speak because we are human. We are more than just language

factories, and language plays just one part in our complex social interactions. Language models, on the other hand, *are* language factories: capable of producing and processing words at astonishing rates, but lacking the faculties that drive human language production and processing.

Linguistics has long seen its subject (language) isolated from all other cognitive (and physical) capabilities. While that enables the study of certain aspects in isolation and develop theories and models, it obscured the larger picture. Sociolinguistics, psycholinguistics, and other sub-fields have fought hard to reintroduce these aspects to the larger linguistic field. NLP today is following a similar trajectory. With language models covering the basis for language production and comprehension, we can return to social aspects of language models. Understanding the social aspects of language technologies requires a refocus on emotional intelligence, cultural factors, values, and norms, as well as social interaction and the broader social implications. In addition, the future of socially aware NLP should emphasize ethical considerations and the responsible development process.

Socially aware NLP will likely transform industries and societal function, as well as shaping the broader field of AI, including but not limited to audio, vision, and robotics, where social awareness can play an even more critical role. For instance, the integration of social awareness in robotics can facilitate the development of robots that can safely and effectively interact with humans (e.g., eldercare robots, service robots) and lead to the advancements in computer vision that enable systems to better interpret emotions (Mittal et al., 2020; Kwon et al., 2023), social interactions, and cultural contexts from visual data (Kruk et al., 2023; Achlioptas et al., 2021). Again, we must proceed with a keen awareness of ethics and risks (Barrett et al., 2023).

In the future, we can also explore how these models work as social agents, which social cues they read and understand, and which tasks requiring social awareness they can master. This will require new tasks, metrics, and approaches, which are strikingly different from the goals we have followed as a field thus far. Most of all, it will require a re-alignment of the currently fractured AI landscape: we will need to work across fields to integrate models of emotions, values, and cultures into the models we have. There are plenty of unexplored research areas waiting to be explored.

8

7 Limitations

Our work focuses on the call for socially aware language technologies and provides an overview of what socially aware NLP is and how to build it. In spite of providing a general framework, our work does not provide any empirical evidence in the form of performance comparisons or other forms of quantitative measures. Additionally, socially aware NLP is linked to a variety of research fields, such as sociolinguistics, semantics, pragmatics, human-computer interaction, and social sciences; our current work only discuss the differences between socially aware NLP and a few very similar directions like computational social science and human-centered NLP. Our position paper also focuses only on a few key concepts related to socially aware NLP; nevertheless, there are many keywords and related subfields, from trust to equity in socio-technical systems, to transparency and societal values. Furthermore, we have discussed different dimensions of social awareness especially around social factors, social interaction and social implications, but it still remains challenging to computationally categorize, evaluate and even visualize social awareness in a system. Finally, we provide a general framework for defining socially aware NLP, but not specific technical methodologies to enable social awareness across different domains, as this would require diverse design choices, which we leave to future work.

8 Ethical Considerations

Developing socially aware NLP systems might require further information from users, and sometimes users may not be fully aware of and consent to such interactions, resulting in increased privacy risks. There is a risk that socially aware, more responsive language technologies can be misused to manipulate and personalize content, spread misinformation, or even persuade people towards certain decisions and behaviors. It is very likely that socially aware NLP systems will become out of date over time or in other contexts and should be continuously monitored in order to ensure that they meet users' varied needs in order to function safely and fairly. NLP systems that are socially aware also pose a risk of users' over-reliance, which can degrade human abilities and skills. Thus, we argue that, building socially aware NLP requires serious ethical considerations and a responsible development process, as addressing biases, ensuring privacy, and taking into account potential misuse becomes increasingly important in building more transparent and accountable language technologies.

References

Gavin Abercrombie, 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, Singapore. Association for Computational Linguistics.

Panos Achlioptas, 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.

Gati V Aher, Rosa I Arriaga, and Adam Tauman Kalai. 2023. Using large language models to simulate multiple humans and replicate human subject studies. In *International Conference on Machine Learning*, pages 337–371. PMLR.

Alaa Alharbi and Mark Lee. 2022. Classifying arabic crisis tweets using data selection and pre-trained language models. In *Proceedinsg of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools with Shared Tasks on Qur'an QA and Fine-Grained Hate Speech Detection*, pages 71–78.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Clark Barrett, 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.

Federico Bianchi, 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*.

Steven Bird 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, St. Julian's, Malta. Association for Computational Linguistics.

Su Lin Blodgett, 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, Online. Association for Computational Linguistics.

Tolga Bolukbasi, 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.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.

Cynthia Breazeal. 2003. Emotion and sociable humanoid robots. *International journal of human-computer studies*, 59(1-2):119–155.

Luke Breitfeller, Emily Ahn, David Jurgens, and Yulia Tsvetkov. 2019. Finding microaggressions in the wild: A case for locating elusive phenomena in social media posts. 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 1664–1674.

Tom Brown, 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.

Tuhin Chakrabarty, Philippe Laban, Divyansh Agarwal, Smaranda Muresan, and Chien-Sheng Wu. 2024. Art or artifice? large language models and the false promise of creativity. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–34.

Lan Chen, Xi Chen, Shiyu Wu, Yaqi Yang, Meng Chang, and Hengshu Zhu. 2023. The future of chatgpt-enabled labor market: A preliminary study. *arXiv preprint arXiv:2304.09823*.

Minje Choi, 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.

Minje Choi, 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.

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

Sunipa Dev, 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.

Shehzaad Dhuliawala, 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, Singapore. Association for Computational Linguistics.

Emily Dinan, 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, Dublin, Ireland. Association for Computational Linguistics.

Esin Durmus, Karina Nyugen, Thomas I Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, et al. 2023. Towards measuring the representation of subjective global opinions in language models. *arXiv preprint arXiv:2306.16388*.

Tyna Eloundou, 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*.

Mai ElSherief, 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.

Alircza Fathi, 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. IEEE.

Lucie Flek. 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, Online. Association for Computational Linguistics.

Vinitha Gadiraju, Shaun Kane, Sunipa Dev, Alex Taylor, Ding Wang, Emily Denton, and Robin Brewer. 2023. " i wouldn't say offensive but...": Disability-centered perspectives on large language models. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 205–216.

Aparna Garimella, Carmen Banea, Dirk Hovy, and Rada Mihalcea. 2019. Women's syntactic resilience and men's grammatical luck: Gender-bias in part-of-speech tagging and dependency parsing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3493–3498.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. 2020. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. *arXiv preprint arXiv:2009.11462*.

Vinicius G Goecks and Nicholas R Waytowich. 2023. Disasterresponsegpt: Large language models for accelerated plan of action development in disaster response scenarios. *arXiv preprint arXiv:2306.17271*.

Hila Gonen 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, Minneapolis, Minnesota. Association for Computational Linguistics.

Erin Grant, Aida Nematzadeh, and Thomas L Griffiths. 2017. How can memory-augmented neural networks pass a false-belief task? In *CogSci*.

Anhong Guo, 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.

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

Gina Hernez-Broome. 2012. Social intelligence: the new science of human relationships.

Daniel Hershcovich, 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, Dublin, Ireland. Association for Computational Linguistics.

Dirk Hovy. 2015. Demographic factors improve classification performance. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 752–762, Beijing, China. Association for Computational Linguistics.

Dirk Hovy, 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, Online. Association for Computational Linguistics.

Dirk Hovy 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, Berlin, Germany. Association for Computational Linguistics.

Dirk Hovy 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.

Jing Huang and Diyi Yang. 2023. Culturally aware natural language inference. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7591–7609.

Muhammad Imran, Prasenjit Mitra, and Carlos Castillo. 2016. Twitter as a lifeline: Human-annotated twitter corpora for nlp of crisis-related messages. *arXiv preprint arXiv:1605.05894*.

Mohit Iyyer, 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.

Maurice Jakesch, 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.

Zhijing Jin, 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.

Pratik Joshi, 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, Online. Association for Computational Linguistics.

Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. 2023. Chatgpt for good? on opportunities and challenges of large language models

for education. *Learning and individual differences*, 103:102274.

Bhushan Kotnis, Kiril Gashteovski, Julia Gastinger, Giuseppe Serra, Francesco Alesiani, Timo Sztyler, Ammar Shaker, Na Gong, Carolin Lawrence, and Zhao Xu. 2022. Human-centric research for nlp: Towards a definition and guiding questions.

Austin C Kozlowski, Matt Taddy, and James A Evans. 2018. The geometry of culture: Analyzing meaning through word embeddings. *arXiv preprint arXiv:1803.09288*.

Julia Kruk, 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.

Minae Kwon, Hengyuan Hu, Vivek Myers, Siddharth Karamcheti, Anca Dragan, and Dorsa Sadigh. 2023. Toward grounded social reasoning. *arXiv preprint arXiv:2306.08651*.

Maria Knight Lapinski and Rajiv N Rimal. 2005. An explication of social norms. *Communication theory*, 15(2):127–147.

Anne Lauscher, Federico Bianchi, Samuel Bowman, and Dirk Hovy. 2022. Socioprobe: What, when, and where language models learn about sociodemographics. *arXiv preprint arXiv:2211.04281*.

David Lazer, Alex Pentland, Lada Adamic, Sinan Aral, Albert-László Barabási, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, et al. 2009. Computational social science. *Science*, 323(5915):721–723.

Matthew Le, 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.

Belinda Z. Li, Alex Tamkin, Noah Goodman, and Jacob Andreas. 2023. Eliciting human preferences with language models.

Robert Litschko, 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*.

Ruibo Liu, Ruixin Yang, Chenyan Jia, Ge Zhang, Diyi Yang, and Soroush Vosoughi. 2023. Training socially aligned language models on simulated social interactions. In *The Twelfth International Conference on Learning Representations*.

Ruibo Liu, Ge Zhang, Xinyu Feng, and Soroush Vosoughi. 2022a. Aligning generative language models with human values. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 241–252.

Yihe Liu, Anushk Mittal, Diyi Yang, and Amy Bruckman. 2022b. 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.

Weicheng Ma, Henry Scheible, Brian Wang, Goutham Veeramachaneni, Pratim Chowdhary, Alan Sun, Andrew Koulogeorge, Lili Wang, Diyi Yang, and Soroush Vosoughi. 2023. Deciphering stereotypes in pre-trained language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11328–11345.

John McCarthy, 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.

Trisha Mittal, 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.

Philipp Mondorf and Barbara Plank. 2024. Comparing inferential strategies of humans and large language models in deductive reasoning.

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

Dong Nguyen, 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.

Ashwin Paranjape and Christopher D Manning. 2021. Human-like informative conversations: Better acknowledgements using conditional mutual information. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 768–781.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–22.

Alex Pentland. 2005. Socially aware computation and communication. In *Proceedings of the 7th international conference on Multimodal interfaces*, pages 199–199.

12

Ildikó Pilán, Laurent Prévot, Hendrik Buschmeier, and Pierre Lison. 2023. Conversational feedback in scripted versus spontaneous dialogues: A comparative analysis. *arXiv preprint arXiv:2309.15656*.

Steven Pinker. 2003. *The language instinct: How the mind creates language*. Penguin uK.

Kim Plunkett. 1997. Theories of early language acquisition. *Trends in cognitive sciences*, 1(4):146–153.

Vinodkumar Prabhakaran, 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, Nagoya, Japan. Asian Federation of Natural Language Processing.

David Premack and Guy Woodruff. 1978. Does the chimpanzee have a theory of mind? *Behavioral and brain sciences*, 1(4):515–526.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain yourself! leveraging language models for commonsense reasoning. *arXiv preprint arXiv:1906.02361*.

Surangika Ranathunga 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*.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy. Association for Computational Linguistics.

Maarten Sap, 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, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Deven Santosh Shah, 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, Online. Association for Computational Linguistics.

Omar Shaikh, Valentino Chai, Michele J Gelfand, Diyi Yang, and Michael S Bernstein. 2023a. Rehearsal: Simulating conflict to teach conflict resolution. *arXiv preprint arXiv:2309.12309*.

Omar Shaikh, Kristina Gligorić, Ashna Khetan, Matthias Gerstgrasser, Diyi Yang, and Dan Jurafsky. 2023b. Grounding or guesswork? large language models are presumptive grounders. *arXiv preprint arXiv:2311.09144*.

Ashish Sharma, 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*.

Bangzhao Shu, 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*.

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

Yueqi Song, 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*.

Taylor Sorensen, 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*.

Christian Stab and Iryna Gurevych. 2014. Identifying argumentative discourse structures in persuasive essays. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 46–56.

Simone Tedeschi, 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, Toronto, Canada. Association for Computational Linguistics.

Michael Tomasello. 2014. *A natural history of human thinking*. Harvard University Press.

Alan M. Turing. 1950. I.—COMPUTING MACHINERY AND INTELLIGENCE. *Mind*, LIX(236):433–460.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting.

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

13

on Empirical Methods in Natural Language Processing, pages 3003–3008, Brussels, Belgium. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Rose E. Wang, 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*. Association for Computational Linguistics.

Leander Weber, Sebastian Lapuschkin, Alexander Binder, and Wojciech Samek. 2023. Beyond explaining: Opportunities and challenges of xai-based model improvement. *Information Fusion*, 92:154–176.

Yuwei Wu, Xuezhe Ma, and Diyi Yang. 2021. Personalized Response Generation via Generative Split Memory Network. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Diyi Yang, Jiaao Chen, Zichao Yang, Dan Jurafsky, and Eduard Hovy. 2019a. Let's make your request more persuasive: Modeling persuasive strategies via semi-supervised neural nets on crowdfunding platforms. 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 3620–3630, Minneapolis, Minnesota. Association for Computational Linguistics.

Diyi Yang, Robert E Kraut, Tenbroeck Smith, Elijah Mayfield, and Dan Jurafsky. 2019b. Seekers, providers, welcomers, and storytellers: Modeling social roles in online health communities. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–14.

Kayo Yin, Amit Moryossef, Julie Hochgesang, Yoav Goldberg, and Malihe Alikhani. 2021. Including signed languages in natural language processing. *arXiv preprint arXiv:2105.05222*.

Ann Yuan, Andy Coenen, Emily Reif, and Daphne Ippolito. 2022. Wordcraft: story writing with large language models. In *27th International Conference on Intelligent User Interfaces*, pages 841–852.

Caleb Ziems, Jiaao Chen, Camille Harris, Jessica Anderson, and Diyi Yang. 2022a. Value: Understanding dialect disparity in nlu. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3701–3720.

Caleb Ziems, William Held, Jingfeng Yang, and Diyi Yang. 2022b. Multi-value: A framework for cross-dialectal english nlp. *arXiv preprint arXiv:2212.08011*.