Advancing Social Intelligence in AI Agents: Technical Challenges and Open Questions

Leena Mathur[1], Paul Pu Liang[2], Louis-Philippe Morency[1]
Language Technologies Institute[1], Machine Learning Department[2]
School of Computer Science, Carnegie Mellon University
{lmathur,pliang,morency}@cs.cmu.edu

Abstract

Building socially-intelligent AI agents (Social-AI) is a multidisciplinary, multimodal research goal that involves creating agents that can sense, perceive, reason about, learn from, and respond to affect, behavior, and cognition of other agents (human or artificial). Progress towards Social-AI has accelerated in the past decade across several computing communities, including natural language processing, machine learning, robotics, human-machine interaction, computer vision, and speech. Natural language processing, in particular, has been prominent in Social-AI research, as language plays a key role in constructing the social world. In this position paper, we identify a set of underlying technical challenges and open questions for researchers across computing communities to advance Social-AI. We anchor our discussion in the context of social intelligence concepts and prior progress in Social-AI research.

1 Introduction

Humans rely on *social intelligence* to interpret and respond to social phenomena such as empathy, rapport, collaboration, and group dynamics. Social intelligence competencies that evolved over thousands of years in *Homo sapiens* are hypothesized to have been core factors shaping human cognition and driving the emergence of language, culture, and societies (Wilson, 2012; Emery et al., 2007; Knight et al., 2000; Goody, 1995; Sterelny, 2007). Humans today continually navigate diverse social contexts, from short-term dyadic conversations to long-term relationships. Virtual and embodied AI agents must have social intelligence competencies in order to function seamlessly alongside humans and other AI agents. The complexity of this vision and a set of core technical challenges for developing these agents are visualized in Figure 1.

Building socially-intelligent AI agents (Social-AI) involves developing computational foundations for agents that can sense, perceive, reason about, learn from, and respond to affective, behavioral, and cognitive constructs of other agents (human or artificial). Social-AI research interest has accelerated across computing communities in recent years, including natural language processing (NLP), machine learning (ML), robotics, human-machine interaction, computer vision, and speech (Figure 2). We see Social-AI beginning to support humans in real-world contexts. Virtual text agents have stimulated empathic conversations between humans in online chatrooms (Sharma et al., 2023), and affective signals from wearables have supported well-being (Sano, 2016; Park et al., 2020). Embodied social robots have supported geriatric care (González-González et al., 2021; Fleming et al., 2003), motivated stroke patients (Matarić et al., 2007; Feingold Polak and Tzedek, 2020), assisted youth with autism spectrum condition (Hurst et al., 2020; Scassellati et al., 2012), improved student mental health (Jeong et al., 2020), and collaborated with humans in manufacturing (Sauppé and Mutlu, 2015), among other prosocial applications[1].

Is Social-AI purely an application of AI to social contexts or do underlying technical challenges emerge that are particularly relevant to Social-AI? We believe there are core technical challenges that must be addressed to advance the multidisciplinary, multimodal goal of building Social-AI. Our position paper is driven by the following question: *What are core technical challenges, open questions, and opportunities for researchers to advance Social-AI?* We anchor our position in the context of social intelligence concepts, reviewed in Section 2, and progress in Social-AI, summarized in Section 3. In Section 4, we present 4 core technical challenges, along with opportunities and open questions to advance Social-AI research.

[1]While Social-AI has prosocial applications, it must be advanced in ethical ways. We discuss Social-AI ethical considerations in Section 7 and Appendix B.

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 20541–20560
November 12-16, 2024 ©2024 Association for Computational Linguistics

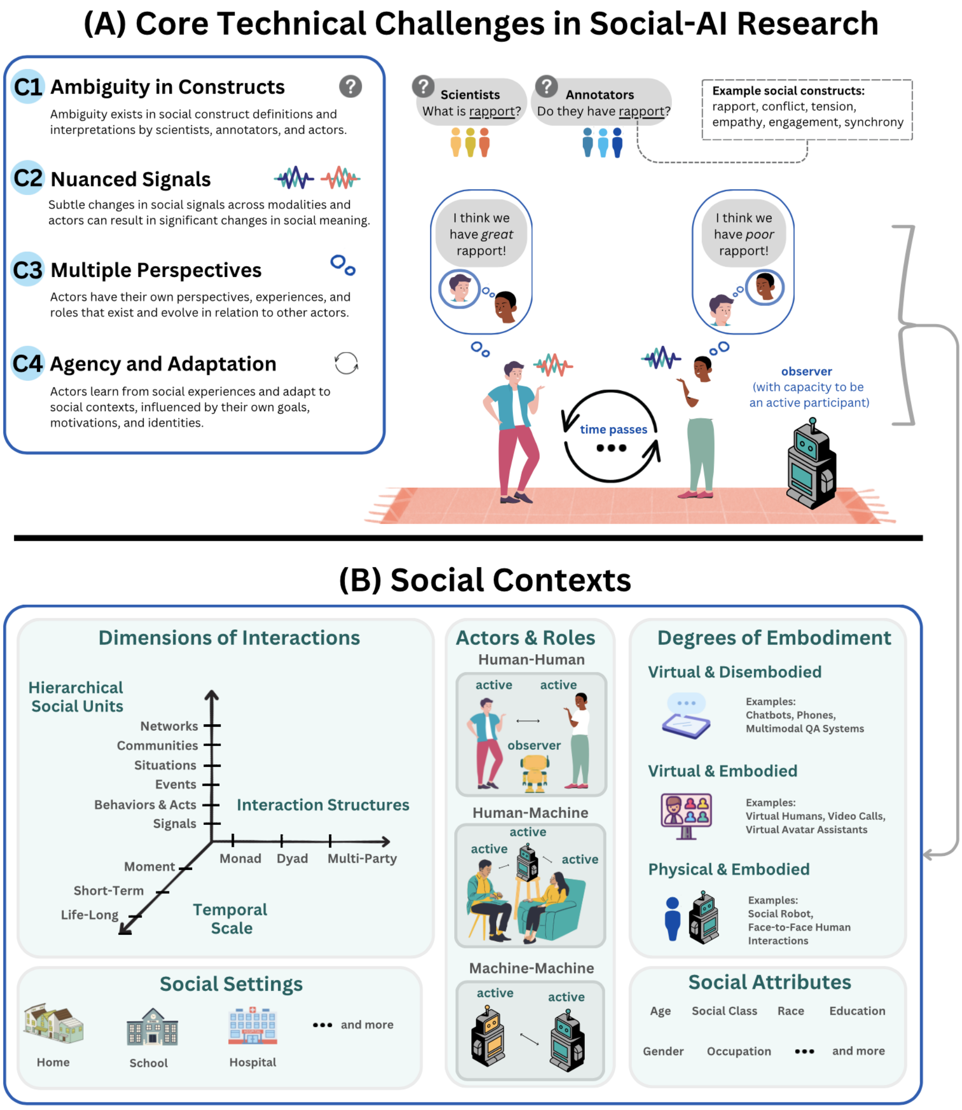

Figure 1: (A) Four core technical challenges in Social-AI research, illustrated in an example context of a Social-AI agent observing and learning from a human-human interaction. (B) Social contexts in which Social-AI agents can be situated, with interactions spanning social units, interaction structures, and timescales. Interactions can span social settings, degrees of agent embodiment, and social attributes of humans, with agents in several roles.

2 Social Intelligence Concepts

A shared understanding of social intelligence concepts is useful for researchers to contextualize progress and challenges in Social-AI. We begin with the concept of *social constructs* – what makes entities social? Social ontologists distinguish between *social constructs* (entities that exist by human construction) and *natural kinds* (entities that exist regardless of the interpretations of human minds) (Searle, 1995, 1998, 2010; Khalidi, 2015). For example, a person is a natural kind with physical properties, but a *friend* or *stranger* are social constructs unlinked to those physical properties. Language plays a key role in forming social constructs (e.g., referring to someone as a *close friend* is an act that can make them a close friend) (Searle, 2012). Social constructs are *ontologically subjective*, as their existence depends on perceptions of humans, referred to as being "perceiver-dependent" (Searle, 1998). This ontological subjectivity of constructs informs the Social-AI challenges that we present in Figure 1A and discuss in Section 4.

2.1 Social Intelligence

The term *social intelligence* was introduced in the early 20th century by social scientists who recognized the importance of teaching children how to comprehend social situations (Dewey, 1909; Lull, 1911). Hypotheses that social intelligence differs from other forms of intelligence began in the 1920s, with distinctions among abstract intelligence (idea-focused), mechanical intelligence (object-focused), and social intelligence (people-focused) (Thorndike, 1920; Thorndike and Stein, 1937), with social intelligence seen as context-dependent (Strang, 1930). Studying properties of social intelligence is an ongoing research area (Conzelmann et al., 2013; Brown et al., 2019).

Humans generate and interpret social signals through verbal and nonverbal sensory channels (e.g., word choice, gaze, gesture) (Vinciarelli et al., 2011; Vinciarelli and Esposito, 2018; Burgoon et al., 2011). Humans perceive social meaning from these signals (Poggi and D'Errico, 2012). A prevalent model developed with construct validation (Weis, 2008; Weis and Süß, 2005; Conzelmann et al., 2013) proposes that social intelligence encompasses 5 competencies enabling humans to navigate social situations: *social perception*, *social knowledge*, *social memory*, *social reasoning*, and *social creativity*. We consider a 6th competency, *social interaction*. Social-AI research attempts to endow agents with these 6 competencies as key elements of social intelligence, defined below.

Social perception involves the ability to perceive socially-relevant information from sensory stimuli (Zebrowitz, 1990; Adolphs et al., 2016) (e.g., reading tension from body language). *Social knowledge* includes both *declarative* (factual) and *procedural* (norms) knowledge (Snyder and Cantor, 1980; Bye and Jussim, 1993). *Social memory* involves the ability to store and recall social knowledge about the self and others (Skowronski et al., 1991; Nelson, 2003). *Social reasoning* involves the ability to interpret social stimuli and make inferences based on these stimuli and commonsense understanding (Gagnon-St-Pierre et al., 2021; Read et al., 2013). *Social creativity*, sometimes referred to as "theory-of-mind", involves the ability to counterfactually reason about social situations (Hughes and Devine, 2015; Astington and Jenkins, 1995). *Social interaction* involves the ability to engage with other agents in mutually co-regulated patterns (Turner, 1988; McCall, 2003; De Jaegher et al., 2010).

2.2 Dimensions of Social Context

Social contexts in which AI agents can be studied and deployed are visualized in Figure 1B. These encompass diverse **social settings** (e.g., homes, hospitals) that influence social norms (e.g., silence in a library, yelling at a hockey match) (Rachlinski, 1998; Axelrod, 1997). Within social contexts, **actors** (human and machine) can have different **roles** (e.g., active participant, observer) and different **social attributes** (e.g, age, occupation) which can shape interactions (Van Bavel et al., 2013; Trepte, 2013; Allen, 2023; Goffman, 2016). The degree of **embodiment** of actors can enrich and augment communication channels (Goodwin, 2000; Wainer et al., 2006; Deng et al., 2019); embodiment spans disembodied virtual agents (e.g., chatbot), embodied virtual agents (e.g., avatar), and physically-embodied agents (e.g., humans and robots).

Within social contexts, there are several **dimensions of interactions**: hierarchical social units, interaction structures, and temporal scale. Social units influence communication content, norms, and interpretation (Hymes et al., 1972; Angelelli, 2000; Agha, 2006; Goody, 1995; Goffman et al., 2002). To visualize social units, consider the following scenario adapted from Hymes et al. (1972): two people exchange eye contact (*social signals*), while one of them tells a joke (*act*) during a conversation (*event*) at a party (*situation*) governed by shared norms (*community*). The dyad's interaction may be viewed as two connected nodes in a *social network* of interactions. Monads, dyads, and multi-party interaction structures can induce their own social dynamics (Pickering and Garrod, 2021). For example, when a third actor is added to a dyad, new dynamics of group coordination and conflict can arise (Olekalns et al., 2003). Social interactions also have a temporal scale that spans split-second communication, moments, short-term interactions, longer-term interactions, and life-long relationships (Wittmann, 2011). This temporal dimension can influence social meaning conveyed during interactions; for example, a 100 millisecond pause longer than normal for an actor can indicate reluctance, instead of eagerness (Durantin et al., 2017).

This paper proposes core technical challenges that are relevant to Social-AI agents situated across many possible dimensions of social context.

3 Progress in Social-AI Research

We examined Social-AI research progress in 6 computing communities: NLP, ML, robotics, human-machine interaction[2], computer vision, and speech. Interest in Social-AI research has accelerated in the past decade, notably in NLP, ML, and robotics (Figure 2). We synthesize key trends to provide readers

[2] We use "human-machine interaction" to include both human-computer interaction and human-robot interaction.

with context for the discussion of core technical challenges and open questions in Section 4. Using the Semantic Scholar API (Kinney et al., 2023), we found 3,257 relevant papers that spanned 1979-2023. For insights, we examined representative papers across communities, selected for citation count, recency, and relevance. Details on search queries and filtering criteria are in Appendix A.

Early Social-AI research primarily envisioned *rule-based* approaches for building social intelligence competencies in agents. NLP and human-machine interaction researchers examined rule-based approaches for modeling goal-oriented communication (Pershina, 1986), cooperation (d'Inverno et al., 1997), and dialogue (McRoy and Hirst, 1993). They also explored rule-based approaches for processing multimodal signals during human-machine interactions (Nagao and Takeuchi, 1994), animating embodied virtual humans as conversational agents (Pelachaud et al., 1991; Cassell et al., 1994; Pelachaud et al., 1996), and interpreting potential effects of speech acts (Pautler and Quilici, 1998). ML and robotics researchers proposed multi-agent search and planning algorithms for social learning in groups of disembodied agents (Ephrati et al., 1993) and *behavior-based* control for groups of robots (Mataric, 1993, 1994). Social robotics advanced systems for attending to social stimuli (e.g., faces) and communicating intent (Breazeal and Scassellati, 1999a,b).

In the past two decades, scientists have increasingly leveraged *ML* and *deep learning* in Social-AI research. A common approach in these methods is to train models to predict social phenomena from observable human behavior, by using *static* datasets with ground truth labels computed as aggregations of annotator perspectives. A focus on ML for *social signal processing* also emerged during this era (Vinciarelli et al., 2009). This stimulated a focus on predicting *social signals*, such as laughter from speech (Brueckner and Schulter, 2014; Eyben et al., 2015), gesture and gait from visual cues (Morency et al., 2007; Chao et al., 2019), and engagement from visual, speech, and physiological data during human-robot interactions (Rudovic et al., 2018). There has been substantial research effort in predicting *affective information*, such as emotion and sentiment, from multimodal conversation signals (Busso et al., 2008; Schuller et al., 2012; Morency et al., 2011; Zadeh et al., 2018; Majumder et al., 2019), as well as efforts to predict social behaviors with affective information (Mathur et al., 2023b).

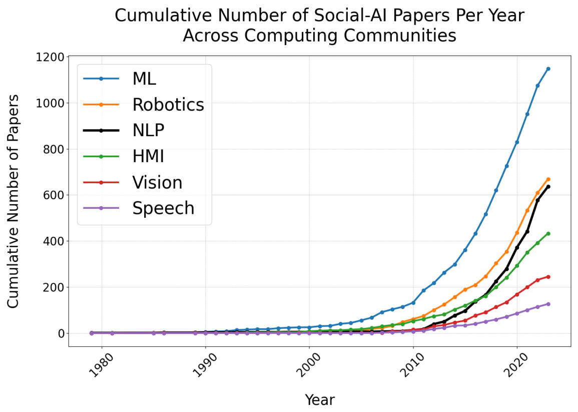

Figure 2: Cumulative number of Social-AI papers over time, based on the 3,257 papers from our Semantic Scholar Social-AI queries. Interest in Social-AI research has been accelerating across computing communities.

Multimodal interaction and computer vision research explored approaches for rendering *virtual humans* with social behavior (Swartout et al., 2006; Ng et al., 2022) and learning representations to model social interactions (Soleymani et al., 2019; Lee et al., 2024). Computer vision and robotics researchers have studied ML approaches for predicting human *intent* to inform human-robot interactions (Strabala et al., 2012) and *social navigation* to improve robot navigation in spaces with humans (Pellegrini et al., 2010; Kosaraju et al., 2019; Cuan et al., 2022; Taylor et al., 2022).

In parallel, researchers have focused on integrating *game-theoretic* and *probabilistic* approaches to model aspects of social intelligence, such as cooperation, competition, and theory-of-mind (Castelfranchi, 1998; Kleiman-Weiner et al., 2016). Approaches include stochastic games, inverse planning, and multi-agent reinforcement learning (MARL) (Jaques et al., 2019; Shum et al., 2019; Wu et al., 2020), studied through formalized games, such as public goods games (Wang et al., 2023), or grid-world simulations (Lee et al., 2021).

In recent years, there have been efforts to probe the extent to which models, in particular large language models (LLMs), exhibit social intelligence competencies. Scientists have identified strong performance by LLMs in procedurally-generated multi-agent gridworld interaction tasks (Kovač et al., 2023), as well as limitations in LLM *social knowledge* and *social reasoning* competencies (Sap et al., 2022; Shapira et al., 2023) when tested on *static* text benchmarks such as SocialIQa (Sap et al., 2019b), ToMI (Le et al., 2019), and Epistemic Reasoning (Cohen, 2021). These model limitations also emerge in *static* VideoQA benchmarks such as SOCIAL-IQ 1.0 (Zadeh et al., 2019), SOCIAL-IQ 2.0 (Wilf et al., 2023), and synthetic video benchmarks such as MMTOM-QA (Jin et al., 2024). There have been efforts to study LLM *social interaction* competency during goal-oriented dialogue (Hosseini-Asl et al., 2020) and open-domain dialogue (Nedelchev et al., 2020). Recently, *dynamic* environments have been proposed to study Social-AI agents in dyadic and multi-party settings, with SOTOPIA (Zhou et al., 2023), CAMEL (Li et al., 2023), and AGENTVERSE (Chen et al., 2023) studying text agent interactions, COELA (Zhang et al., 2023b) studying embodied language agent interactions, and HABITAT 3.0 (Puig et al., 2023) studying simulations of human-robot interactions.

Key Takeaways Social-AI research has made substantial progress in modeling social phenomena in *static* ungrounded data (abstracting away physical and social context), synthetic data, and lab-controlled social interactions. However, these data can abstract away the *richness* of multimodality (Liang et al., 2024) in interactions, as well as the *context-sensitivity* and *ambiguity* inherent to social phenomena in-the-wild. We also found a focus on modeling *temporally-localized* phenomena (e.g., split-second communication, short-term interactions); longer-term phenomena were comparatively understudied (e.g., hours, days, years).

4 Core Technical Challenges and Open Questions

Social intelligence concepts (Section 2) and prior research in Social-AI[3] (Section 3) informed our identification of 4 core technical challenges visualized in Figure 1A: **(C1) ambiguity in constructs, (C2) nuanced signals, (C3) multiple perspectives, (C4) agency and adaptation**. We believe these challenges are particularly relevant to Social-AI and must be addressed to advance social intelligence in AI agents situated across the landscape of social contexts in Figure 1B[4]. In this section, we present these challenges, along with opportunities and open questions for Social-AI research.

[3]We contribute a repository of resources to inform researchers addressing challenges, discussed in Appendix B and linked here: `https://github.com/l-mathur/social-ai`

[4]We sought to disentangle *scale* (e.g., number of actors and constructs) from the identification of core challenges.

4.1 (C1) Ambiguity in Constructs

Social constructs have inherent ambiguity in their definition and interpretation in the social world. For example, consider the interaction in Figure 1A – how might we characterize rapport, conflict, or tension between the actors? Many of these social constructs are still being defined and operationalized by scientists (Neequaye, 2023; Policarpo, 2015). The ontological subjectivity of social constructs (entities that exist by human construction, as discussed in Section 2) results in inherent **ambiguity in construct definitions**. This ambiguity is amplified when defining and measuring **hierarchical constructs** – social constructs composed of other social constructs. For example, it has been theorized that rapport between humans can be measured by composing estimates of mutual attentiveness, positivity, and coordination (Tickle-Degnen and Rosenthal, 1990). How might we interpret and measure these inherently ambiguous components? This is compounded when quantifying observations of a social construct ("some rapport", "a little conflict", "a lot of tension"). When modeling ambiguous social constructs, there is likely to also be misalignment in interpretations of these constructs by actors within interactions and annotators viewing interactions. This misalignment amplifies **ambiguity in ground truth** of social constructs. For example, in Figure 1A, it is challenging to assign a label that conclusively represents rapport in the interaction.

C1 Opportunities and Open Questions Researchers must reconsider methods for representing ambiguous social constructs in modeling approaches. When there exists ambiguity in social construct ground truth, as observed in annotator ratings (Yannakakis and Martínez, 2015), how might we incorporate this ambiguity in Social-AI modeling approaches? Modeling efforts have largely relied on predefined label sets and "gold standard" annotations that aggregate annotators' interpretations of social constructs into discrete labels (computed by majority vote) or continuous labels (computed by temporal alignment) (Nicolaou et al., 2014; Kossaifi et al., 2019). However, annotators' and actors' definitions and interpretations of ambiguous social constructs can vary during the course of interactions. We believe such social constructs cannot be solely represented with a single discrete or continuous number on a *static* scale for an interaction.

Natural language, with its ability to expressively represent concepts (Liu and Singh, 2004) and construct the social world (Searle, 1995), may offer a uniquely effective modality to help address sources of ambiguity when representing and modeling social constructs. We see considerable opportunity for Social-AI researchers to explore techniques for leveraging richer natural language descriptions of social phenomena in label spaces. Researchers can explore approaches for constructing flexible natural language label spaces that are *dynamically-generated* and *adjusted* during training and inference, instead of relying on predefined static labels to categorize inputs. These flexible label spaces would enable the development of models for social constructs that can represent greater ambiguity in social contexts. How to best design frameworks to accommodate flexible and dynamic label spaces when modeling social constructs remains an open question and research opportunity.

We direct readers to representative work in areas that can inform Social-AI research towards this challenge: questioning the assumption that labels in ML tasks must have one "ground truth" (Aroyo and Welty, 2015; Cabitza et al., 2023), exploring ordinal representations (e.g., relative scales) for emotion annotation (Yannakakis et al., 2017), modeling perception uncertainty across annotators (Han et al., 2017; Ghandeharioun et al., 2019), inferring annotator subjectivity across samples (Sampath et al., 2022), evaluating the alignment of "ground truth" annotations with diverse annotator perspectives (Santy et al., 2023), and handling varied label distributions (e.g., label distribution learning, multi-label learning) (Geng, 2016; Liu et al., 2021).

4.2 (C2) Nuanced Signals

Social constructs are expressed through behaviors and signals that can be nuanced, often manifesting through different degrees of synchrony across actors and modalities. During interactions, small changes in social signals can lead to large shifts in social meaning being conveyed. For example, an actor's slight change in posture or split-second vocal emphasis on a particular word can communicate rapport. The challenge lies in enabling Social-AI agents to **perceive and generate fine-grained social signals** (e.g., chatbot sensing a user's slowly-building frustration, robot making subtle gestures). Social signals can be expressed with different degrees of **synchrony** and can be **interleaved** across actors and across modalities, with actors functioning as both speakers and listeners (Morency, 2010; Watzlawick et al., 2011). The challenge involves advancing social signal processing techniques (Shmueli et al., 2014) and multimodal models that operate upon verbal and nonverbal information to interpret the nuances of cross-actor and cross-modal interaction patterns.

C2 Opportunities and Open Questions: While scientists are studying the capacity of language to scaffold visual understanding (El Banani et al., 2023; Rozenberszki et al., 2022), audio understanding (Elizalde et al., 2023), and virtual agent motion generation (Zhai et al., 2023; Zhang et al., 2023c), the role of natural language supervision for processing nuanced multimodal social signals remains an open question. We anticipate that advancing the ability of Social-AI systems to process highly-nuanced signals will require researchers to critically examine the role of language in guiding social perception and behavior generation. To what extent can language be treated as an *intermediate representation* (Zeng et al., 2022) to represent and integrate nuanced multimodal social signals? Are there social signals during interactions that cannot be effectively described in language? We expect that processing fine-grained signals such as vocal cues, eye movements (Adams Jr et al., 2017), and gestures (often interleaved within milliseconds), alongside natural language, will require Social-AI researchers to develop new frameworks for representing and aligning social signals across multiple actors and multiple modalities.

Another modeling consideration related to this challenge lies in developing mechanisms for agents to perceive the *absence* of cues. Most ML models are trained to learn representations of attributes based on the *presence* of those attributes in data, yet much nuance in social perception lies in recognizing the absence of stimuli, such as unsaid words, omitted eye contact, silences, and failure to adhere to social norms. Humans are theorized to learn from *absent* cues in addition to visible cues, per error-driven and implicit learning frameworks (Markman, 1989; Van Hamme and Wasserman, 1994; Nixon et al., 2022). Researchers have considerable opportunity to explore approaches for addressing the *absence* of cues in algorithms for agent perception of nuanced social signals.

It remains unknown how abilities in nuanced social perception or social behavior generation might emerge in models during training or fine-tuning (Gururangan et al., 2020; Wei et al., 2022; Arora and Goyal, 2023), leading to the following questions: What are the capabilities and limitations of training objectives, such as next-token prediction or masking, in inducing nuanced social understanding in models? How might the type and amount of training data influence a model's abilities in social understanding? How might tokenization schemes influence a model's abilities in social understanding? While inductive biases of tokenization (Singh and Strouse, 2024) and fine-grained understanding in models have been studied in non-social domains (e.g., segmentation, spatial relationships) (Krause et al., 2015; Bugliarello et al., 2023; Guo et al., 2024), fine-grained social understanding abilities have been understudied. Research in this direction would advance community understanding of how social intelligence competencies might (or might not) emerge in models under various training paradigms. Datasets with nuanced social signals will be essential to pursue these open questions; we discuss data considerations in Appendix B.

4.3 (C3) Multiple Perspectives

In social interactions, actors bring their own perspectives, experiences, and roles; these factors can change over time and influence the perspectives of other actors during interactions. The **subjectivity** in each actor's perceptions of social situations stems from the "perceiver-dependent" nature of social constructs (Section 2). For example, in the interaction in Figure 1A, both actors have different perspectives on the level of "rapport" in their interaction. The evolution of these **concurrent multiple perspectives** can be influenced by several factors of the interaction's social context. Actors meeting for the first time will use different information to estimate rapport than if they had a long-term relationship with frequent interactions to draw upon. The topic of conversation, social setting (e.g., are they in a doctor's office?), behavioral norms (Ziems et al., 2023), social roles (e.g., are they colleagues?) and additional social attributes (e.g., age) can influence the evolution of their concurrent perspectives.

In addition, there exists **multi-perspective interdependence**, as each actor's perspective can change over time, while influencing and being influenced by others' perspectives. For example, a hospital patient and assistive Social-AI agent interacting intermittently would be continually adjusting their perspective of the other actor to build rapport over time. The challenge involves equipping Social-AI agents with the capacity to reason over these multi-perspective dynamics.

C3 Opportunities and Open Questions: Addressing this challenge will require modeling paradigms that enable agents to reason over multiple, dynamically-changing perspectives, experiences, and roles in social interactions. Each actor can influence and be influenced by other actors, as formalized in interactionist theories from psychology such as *actor-partner* frameworks (Kenny and Ledermann, 2010), *social identity theory* (Hogg, 2016), and *social influence theory* (Friedkin and Johnsen, 2011). This complexity leads us to identify the following open questions: How can researchers create models for Social-AI agents to perceive concurrent, interdependent perspectives of actors during interactions? To what extent would a single, joint model be more effective than multiple models (e.g., one for each actor) to represent social phenomena across an interaction? When interactions occur intermittently over time, how can models efficiently and accurately adjust agents' perceptions of other actors' perspectives?

The social intelligence competencies of social creativity (theory-of-mind) and social reasoning (Section 2.1) are typically associated with this "multi-perspective" modeling challenge. Existing theory-of-mind research in Social-AI has focused on the movement of abstract shapes (Gordon, 2016), procedurally-generated dialogues (Kim et al., 2023), and synthetic videos (Jin et al., 2024) and has been useful in assessing abilities of models in these tasks. However, we believe that the multi-perspective Social-AI challenge encompasses not only the capacity to reason about the mental states and beliefs of other agents (including higher-order theory-of-mind), but also the capacity to perform counterfactual social reasoning *over time* with multiple dimensions of dynamic social contexts influencing interdependent perspectives of the actors. In order to test models' ability to address this challenge, the Social-AI community will require the curation of fine-grained benchmarks that test multi-perspective abilities in naturalistic social settings. Data considerations for the Social-AI community are discussed in Appendix B.

4.4 (C4) Agency and Adaptation

Actors learn from social experiences and adapt to social contexts, through interactions, influenced by their own agency, goals, motivations, and identities. Social-AI agents must have the capacity to be *goal-oriented*, often targeting multiple goals simultaneously. For goal-oriented Social-AI agents to **learn from social experiences**, these agents must have mechanisms for motivation to **adapt to both explicit and implicit social signals** from other actors and from the surrounding social context. *Explicit social signals* can include direct or rank-based feedback provided by humans assessing the perceived competencies of an AI agent driven by a social goal. *Implicit signals* can include verbal and nonverbal cues from actors, as well as latent aspects of the social environment (e.g., community-level norms) that can provide supervision for a goal-oriented agent. For example, in Figure 1A, given a social goal of faciliting rapport, the robot might use its observations of the dyad's explicit and implicit social signals to learn how to successfully intervene in the interaction. Creating goal-oriented Social-AI agents will involve developing computational mechanisms for **shared social memory** between a Social-AI agent attempting to learn from social experiences and other actors in an interaction. Building a shared social reality across actors within an interaction (Echterhoff et al., 2009) enables agents to create common ground (Clark and Brennan, 1991) and operate in ways aligned with mutual social expectations. Alignment in these social expectations will be necessary for Social-AI agents to effectively adapt behavior in relation to other actors and achieve long-term social goals.

C4 Opportunities and Open Questions: For many non-social tasks, motivation for agents to learn how to exhibit certain behaviors can be instilled through clearly-defined loss functions and metrics (e.g., minimizing bounding box error for object detection), as well as reward signals based on these metrics. However, Social-AI agents need to learn from multiple kinds of *social experiences* (Hu, 2021). In many cases, this can involve learning from implicit social signals that are *fleeting* (e.g., vocal cue lasting 1 second), *sparse* (e.g., an actor raising an eyebrow once to indicate disapproval), and *context-dependent* (e.g., varied cultural norms). Learning from these types of social signals will be important, as it is infeasible to expect humans to provide regular, explicit feedback to indicate satisfaction with a Social-AI agent's competencies. It is possible that humans will be unaware of the Social-AI agent's goals, reducing the likelihood of humans providing explicit feedback in-the-wild about the agent's performance.

The challenge for researchers is to build mechanisms that motivate Social-AI agents to learn from diverse spaces of social signals. This challenge leads us to identify several understudied, open questions: How can modeling paradigms and metrics be created for Social-AI agents to estimate success in achieving social goals, based on explicit and implicit signals? How can mechanisms to adapt behavior towards achieving *single goals* and *multiple, simultaneous goals* be developed for agents? How might social rewards (Tamir and Hughes, 2018) shape agent learning over time? How can shared social memory be built among Social-AI agents and other actors in interactions, and how can this memory inform algorithms for learning from social signals? To begin addressing these directions, researchers might consider perspectives on motivation (Chentanez et al., 2004), learning from implicit signals (Jaques et al., 2020; Wang et al., 2022), value internalization (Rong and Kleiman-Weiner, 2024) from social feedback, and agent memory (Cheng et al., 2024; Zhong et al., 2024). There exists considerable opportunity to tackle open questions related to Social-AI agency and adaptation.

5 Conclusion

In this position paper, we present a set of core technical challenges, along with opportunities and open questions, for advancing Social-AI research across computing communities. The core technical challenges that we identify include the following: (C1) ambiguity in constructs, (C2) nuanced signals, (C3) multiple perspectives, and (C4) agency and adaptation in social agents. We believe these challenges are relevant for developing Social-AI agents situated in diverse social contexts. For example, NLP researchers situating chatbots within social dialogue and robotics researchers building social robots will both encounter these 4 challenges.

Our research vision for building Social-AI encompasses work to advance social agents with a variety of embodiments, social attributes, and social roles, with agents interacting in a range of social contexts. We conceptualize dimensions of social context and illustrate the core technical challenges in Figure 1, as a resource for readers across fields to visualize a holistic perspective of Social-AI.

We anchor our paper in key perspectives from the social intelligence literature (Section 2) and progress in Social-AI research across multiple computing communities and multiple decades (Section 3). We contribute these summaries to stimulate a shared community understanding on Social-AI.

Virtual and embodied AI systems with social intelligence have considerable potential to support human health and well-being in real-world contexts, such as homes, hospitals, and other shared spaces. For AI agents to function seamlessly alongside humans, these agents need to be endowed with social intelligence competencies. Our paper is intended to inform and motivate research efforts towards AI agents with social intelligence.

6 Limitations

Position Paper Constraints This work is a *position paper*, and the core challenges we identify are informed by our involvement in Social-AI research and our reflection on existing Social-AI papers across NLP, ML, robotics, human-machine interaction, computer vision, and speech. Our search queries and filtering criteria were designed to help us capture and convey trends in Social-AI research from 6 computing communities over multiple decades (Section 3 and Appendix A). We have highlighted representative papers across fields, and this position paper is not an exhaustive survey.

Scope Our paper is scoped to focus on core technical challenges and open questions in Social-AI research, anchored the social intelligence concepts and competencies discussed in Section 2. An important direction of research exists in *social bias* detection and mitigation (Sap et al., 2019a; Blodgett et al., 2020; Liang et al., 2021; Lee et al., 2023). Social-AI research and the deployment of Social-AI systems must be informed by bias and ethics considerations, as discussed in Section 7.

7 Ethics

Social-AI has several prosocial applications to democratize human access to healthcare, education, and other domains (Section 1). Our society faces widening *care gaps* with a shortage of human care providers in mental healthcare (Pathare et al., 2018) and geriatric care (Redfoot et al., 2013), alongside growing global education inequalities (Attewell and Newman, 2010). Social-AI agents have potential to augment human ability to tackle these challenges; for example, through personalized Social-AI agents for education (Park et al., 2019; Spaulding, 2022; Dumont and Ready, 2023). However, while addressing all the Social-AI technical challenges, research must be conducted with ethical and privacy-preserving practices. For example, a scientist advancing the ability of AI systems to detect and generate nuanced social signals (C2) must acknowledge and mitigate the risk of undermining human *trust* in AI systems that are likely to exhibit uncanny behavior in this process (Mori et al., 2012; Mathur and Reichling, 2016).

Participatory Social-AI Research In order to better align development and deployment of *societally-beneficial* Social-AI, we advocate that future directions include embracing Participatory AI frameworks (Bondi et al., 2021; Birhane et al., 2022; Zhang et al., 2023a), consciously involving a diverse range of stakeholders, to ensure Social-AI researchers prioritize concerns, ethical frameworks, risks, and needs raised by stakeholders. When humans interact with computers, virtual agents, and robots, they impose social norms and expectations on these agents (Nass and Moon, 2000); it is important to directly assess and center human social and functional expectations from Social-AI agents when creating and deploying them (Takayama et al., 2008; Dennler et al., 2022; Olatunji et al., 2024). Participatory AI practices will help Social-AI researchers develop guidelines for *transparency* (Felzmann et al., 2020) in how data is gathered, what data is stored, and how data is used through collaboration between end users of Social-AI and researchers of Social-AI. We envision that Participatory Social-AI research would involve advancing *user-centric* modeling paradigms that provide users with power to choose, edit, and delete their data used in models, as well as signal *social* and *cultural* norms they would like their systems to follow.

Social Bias Social-AI systems often rely on models that can exhibit and amplify *social bias*. Techniques to identify and mitigate social bias in language models (Liang et al., 2021; Kaneko and Bollegala, 2022; Prabhumoye et al., 2021) and multimodal models (Luccioni et al., 2024; Cho et al., 2023) are important, in order to build Social-AI systems that are culturally-competent (Bhatt and Diaz, 2024; Bhatia et al., 2024), develop techniques that perform across cultures (Hershcovich et al., 2022; Mathur et al., 2023a), and refrain from propagating harmful social biases when deployed in-the-wild.

Privacy and Trust We believe that Social-AI researchers must prioritize user *privacy*, especially due to the sensitive nature of social phenomena and the deployment of Social-AI agents in spaces alongside humans (Kaminski et al., 2016). Respecting user privacy and building *trust* can involve transparency about how human data is being used, models refraining from tasks they cannot perform (Akter et al., 2024), and minimizing the collection and use of invasive data (Stapels et al., 2023). Models that can learn from minimal data (e.g., a single user's interaction), perform decentralized learning (Guerdan and Gunes, 2022), and store de-identified representations of data can support the privacy of end-users. In addition, developing *on-device* models capable of operating on single devices (e.g., smartphones and wearables) can minimize dependence on external services and reduce risk of exposing sensitive data. These approaches can promote *trustworthiness* of Social-AI systems.

Acknowledgements

This material is based upon work partially supported by National Institutes of Health awards R01MH125740, R01MH132225, R01MH096951 and R21MH130767. Leena Mathur is supported by the NSF Graduate Research Fellowship Program under Grant No. DGE2140739. Paul Pu Liang was supported in part by a Siebel Scholarship and a Waibel Presidential Fellowship. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the sponsors, and no official endorsement should be inferred. Figure 1 includes icon material available from https://icons8.com.

References

Reginald B Adams Jr, Daniel N Albohn, and Kestutis Kveraga. 2017. Social vision: Applying a social-functional approach to face and expression perception. *Current Directions in Psychological Science*, 26(3):243–248.

Ralph Adolphs, Lauri Nummenmaa, Alexander Todorov, and James V Haxby. 2016. Data-driven approaches in the investigation of social perception. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 371(1693):20150367.

Asif Agha. 2006. *Language and social relations*, volume 24. Cambridge University Press.

Syeda Nahida Akter, Sangwu Lee, Yingshan Chang, Yonatan Bisk, and Eric Nyberg. 2024. Visreas: Complex visual reasoning with unanswerable questions. *arXiv preprint arXiv:2403.10534*.

Brenda J Allen. 2023. *Difference matters: Communicating social identity*. Waveland Press.

Claudia Angelelli. 2000. Interpretation as a communicative event: A look through hymes' lenses. *Meta*, 45(4):580–592.

Sanjeev Arora and Anirudh Goyal. 2023. A theory for emergence of complex skills in language models. *arXiv preprint arXiv:2307.15936*.

Lora Aroyo and Chris Welty. 2015. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1):15–24.

Janet Wilde Astington and Jennifer M Jenkins. 1995. Theory of mind development and social understanding. *Cognition & Emotion*, 9(2-3):151–165.

Paul A Attewell and Katherine S Newman. 2010. *Growing gaps: Educational inequality around the world*. Oxford University Press.

Robert Axelrod. 1997. *Promoting Norms*, pages 40–68. Princeton University Press.

Mehar Bhatia, Sahithya Ravi, Aditya Chinchure, Eunjeong Hwang, and Vered Shwartz. 2024. From local concepts to universals: Evaluating the multicultural understanding of vision-language models. *arXiv preprint arXiv:2407.00263*.

Shaily Bhatt and Fernando Diaz. 2024. Extrinsic evaluation of cultural competence in large language models. *arXiv preprint arXiv:2406.11565*.

Abeba Birhane, William Isaac, Vinodkumar Prabhakaran, Mark Diaz, Madeleine Clare Elish, Iason Gabriel, and Shakir Mohamed. 2022. Power to the people? opportunities and challenges for participatory ai. *Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–8.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of" bias" in nlp. *arXiv preprint arXiv:2005.14050*.

Elizabeth Bondi, Lily Xu, Diana Acosta-Navas, and Jackson A Killian. 2021. Envisioning communities: a participatory approach towards ai for social good. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 425–436.

C. Breazeal and B. Scassellati. 1999a. How to build robots that make friends and influence people. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, volume 2, pages 858–863 vol.2.

Cynthia Breazeal and Brian Scassellati. 1999b. A context-dependent attention system for a social robot. *rn*, 255(3).

Matt I Brown, Adrianna Ratajska, Shannon L Hughes, Julie B Fishman, Elisa Huerta, and Christopher F Chabris. 2019. The social shapes test: A new measure of social intelligence, mentalizing, and theory of mind. *Personality and Individual Differences*, 143:107–117.

Raymond Brueckner and Björn Schulter. 2014. Social signal classification using deep blstm recurrent neural networks. In *IEEE International Conference on Acoustics, Speech and Signal processing (ICASSP)*, pages 4823–4827. IEEE.

Emanuele Bugliarello, Laurent Sartran, Aishwarya Agrawal, Lisa Anne Hendricks, and Aida Nematzadeh. 2023. Measuring progress in fine-grained vision-and-language understanding. *arXiv preprint arXiv:2305.07558*.

Judee K Burgoon, Laura K Guerrero, and Valerie Manusov. 2011. Nonverbal signals. *Handbook of Interpersonal Communication*, pages 239–280.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42:335–359.

Lynn Bye and Lee Jussim. 1993. A proposed model for the acquisition of social knowledge and social competence. *Psychology in the Schools*, 30(2):143–161.

Federico Cabitza, Andrea Campagner, and Valerio Basile. 2023. Toward a perspectivist turn in ground truthing for predictive computing. In *AAAI Conference on Artificial Intelligence*, volume 37, pages 6860–6868.

Justine Cassell, Catherine Pelachaud, Norman Badler, Mark Steedman, Brett Achorn, Tripp Becket, Brett Douville, Scott Prevost, and Matthew Stone. 1994. Animated conversation: rule-based generation of facial expression, gesture & spoken intonation for multiple conversational agents. In *Proceedings of the 21st Annual Conference on Computer Graphics and Interactive Techniques*, pages 413–420.

Cristiano Castelfranchi. 1998. Modelling social action for ai agents. *Artificial Intelligence*, 103(1-2):157–182.

Hanqing Chao, Yiwei He, Junping Zhang, and Jianfeng Feng. 2019. Gaitset: Regarding gait as a set for cross-view gait recognition. In *AAAI Conference on Artificial Intelligence*, volume 33, pages 8126–8133.

Weize Chen, Yusheng Su, Jingwei Zuo, Cheng Yang, Chenfei Yuan, Chen Qian, Chi-Min Chan, Yujia Qin, Yaxi Lu, Ruobing Xie, et al. 2023. Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors in agents. *arXiv preprint arXiv:2308.10848*.

Yuheng Cheng, Ceyao Zhang, Zhengwen Zhang, Xiangrui Meng, Sirui Hong, Wenhao Li, Zihao Wang, Zekai Wang, Feng Yin, Junhua Zhao, et al. 2024. Exploring large language model based intelligent agents: Definitions, methods, and prospects. *arXiv preprint arXiv:2401.03428*.

Nuttapong Chentanez, Andrew Barto, and Satinder Singh. 2004. Intrinsically motivated reinforcement learning. *Advances in Neural Information Processing Systems*, 17.

Jaemin Cho, Abhay Zala, and Mohit Bansal. 2023. Dall-eval: Probing the reasoning skills and social biases of text-to-image generation models. In *IEEE/CVF International Conference on Computer Vision*, pages 3043–3054.

Herbert H Clark and Susan E Brennan. 1991. Grounding in communication.

Michael Cohen. 2021. Exploring roberta's theory of mind through textual entailment.

Kristin Conzelmann, Susanne Weis, and Heinz-Martin Süß. 2013. New findings about social intelligence. *Journal of Individual Differences*.

Catie Cuan, Edward Lee, Emre Fisher, Anthony Francis, Leila Takayama, Tingnan Zhang, Alexander Toshev, and Sören Pirk. 2022. Gesture2path: Imitation learning for gesture-aware navigation. *arXiv preprint arXiv:2209.09375*.

Hanne De Jaegher, Ezequiel Di Paolo, and Shaun Gallagher. 2010. Can social interaction constitute social cognition? *Trends in Cognitive Sciences*, 14(10):441–447.

Eric Deng, Bilge Mutlu, Maja J Mataric, et al. 2019. Embodiment in socially interactive robots. *Foundations and Trends in Robotics*, 7(4):251–356.

Nathaniel Dennler, Changxiao Ruan, Jessica Hadiwijoyo, Brenna Chen, Stefanos Nikolaidis, and Maja Mataric. 2022. Using design metaphors to understand user expectations of socially interactive robot embodiments. *arXiv preprint arXiv:2201.10671*.

John Dewey. 1909. *Moral principles in education*. Houghton Mifflin.

Yi Ding, Jacob You, Tonja-Katrin Machulla, Jennifer Jacobs, Pradeep Sen, and Tobias Höllerer. 2022. Impact of annotator demographics on sentiment dataset labeling. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):1–22.

Mark d'Inverno, Michael Luck, and Michael Wooldridge. 1997. Cooperation structures. In *International Joint Conference on Artificial Intelligence*, pages 600–605. Morgan Kaufmann.

Hanna Dumont and Douglas D Ready. 2023. On the promise of personalized learning for educational equity. *Npj Science of Learning*, 8(1):26.

Gautier Durantin, Scott Heath, and Janet Wiles. 2017. Social moments: a perspective on interaction for social robotics. *Frontiers in Robotics and AI*, 4:24.

Gerald Echterhoff, E Tory Higgins, and John M Levine. 2009. Shared reality: Experiencing commonality with others' inner states about the world. *Perspectives on Psychological Science*, 4(5):496–521.

Mohamed El Banani, Karan Desai, and Justin Johnson. 2023. Learning visual representations via language-guided sampling. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19208–19220.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. 2023. Clap learning audio concepts from natural language supervision. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Nathan J Emery, Nicola S Clayton, and Chris D Frith. 2007. Introduction. social intelligence: from brain to culture.

Eithan Ephrati, Jeffrey S Rosenschein, et al. 1993. Multi-agent planning as a dynamic search for social consensus. In *IJCAI*, volume 93, pages 423–429.

Florian Eyben, Klaus R Scherer, Björn W Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y Devillers, Julien Epps, Petri Laukka, Shrikanth S Narayanan, et al. 2015. The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE Transactions on Affective Computing*, 7(2):190–202.

Ronit Feingold Polak and Shelly Levy Tzedek. 2020. Social robot for rehabilitation: expert clinicians and post-stroke patients' evaluation following a long-term intervention. In *ACM/IEEE International Conference on Human-Robot Interaction*, pages 151–160.

Heike Felzmann, Eduard Fosch-Villaronga, Christoph Lutz, and Aurelia Tamò-Larrieux. 2020. Towards transparency by design for artificial intelligence. *Science and Engineering Ethics*, 26(6):3333–3361.

Kevin C Fleming, Jonathan M Evans, and Darryl S Chutka. 2003. Caregiver and clinician shortages in an aging nation. In *Mayo Clinic Proceedings*, volume 78, pages 1026–1040. Elsevier.

Noah E Friedkin and Eugene C Johnsen. 2011. *Social influence network theory: A sociological examination of small group dynamics*, volume 33. Cambridge University Press.

Émilie Gagnon-St-Pierre, Marina M Doucerain, and Henry Markovits. 2021. Reasoning strategies explain individual differences in social reasoning. *Journal of Experimental Psychology: General*, 150(2):340.

Xin Geng. 2016. Label distribution learning. *IEEE Transactions on Knowledge and Data Engineering*, 28(7):1734–1748.

Asma Ghandeharioun, Brian Eoff, Brendan Jou, and Rosalind Picard. 2019. Characterizing sources of uncertainty to proxy calibration and disambiguate annotator and data bias. In *IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 4202–4206. IEEE.

Erving Goffman. 2016. The presentation of self in everyday life. In *Social Theory Re-Wired*, pages 482–493. Routledge.

Erving Goffman et al. 2002. The presentation of self in everyday life. 1959. *Garden City, NY*, 259.

Carina Soledad González-González, Verónica Violant-Holz, and Rosa Maria Gil-Iranzo. 2021. Social robots in hospitals: a systematic review. *Applied Sciences*, 11(13):5976.

Charles Goodwin. 2000. Action and embodiment within situated human interaction. *Journal of Pragmatics*, 32(10):1489–1522.

Esther N Goody. 1995. *Social intelligence and interaction: Expressions and implications of the social bias in human intelligence*. Cambridge University Press.

Andrew Gordon. 2016. Commonsense interpretation of triangle behavior. In *AAAI Conference on Artificial Intelligence*, volume 30.

Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. 2023. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. *arXiv preprint arXiv:2311.18259*.

Luke Guerdan and Hatice Gunes. 2022. Decentralized robot learning for personalization and privacy. *arXiv preprint arXiv:2201.05527*.

Qiushan Guo, Shalini De Mello, Hongxu Yin, Wonmin Byeon, Ka Chun Cheung, Yizhou Yu, Ping Luo, and Sifei Liu. 2024. Regiongpt: Towards region understanding vision language model. *arXiv preprint arXiv:2403.02330*.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Jing Han, Zixing Zhang, Maximilian Schmitt, Maja Pantic, and Björn Schuller. 2017. From hard to soft: Towards more human-like emotion recognition by modelling the perception uncertainty. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 890–897.

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

Michael A Hogg. 2016. *Social identity theory*. Springer.

Ehsan Hosseini-Asl, Bryan McCann, Chien-Sheng Wu, Semih Yavuz, and Richard Socher. 2020. A simple language model for task-oriented dialogue. *Advances in Neural Information Processing Systems*, 33:20179–20191.

Zhiting Hu. 2021. *Towards Training AI Agents with All Types of Experiences: A Standardized ML Formalism*. Ph.D. thesis, Carnegie Mellon University.

Olivia Huang, Eve Fleisig, and Dan Klein. 2023. Incorporating worker perspectives into mturk annotation practices for nlp. *arXiv preprint arXiv:2311.02802*.

Claire Hughes and Rory T Devine. 2015. A social perspective on theory of mind. *Handbook of Child Psychology and Developmental Science*, pages 1–46.

Nikki Hurst, Caitlyn Clabaugh, Rachel Baynes, Jeff Cohn, Donna Mitroff, and Stefan Scherer. 2020. Social and emotional skills training with embodied moxie. *arXiv preprint arXiv:2004.12962*.

Dell Hymes et al. 1972. Models of the interaction of language and social life.

Natasha Jaques, Angeliki Lazaridou, Edward Hughes, Caglar Gulcehre, Pedro Ortega, DJ Strouse, Joel Z Leibo, and Nando De Freitas. 2019. Social influence as intrinsic motivation for multi-agent deep reinforcement learning. In *International Conference on Machine Learning*, pages 3040–3049. PMLR.

Natasha Jaques, Judy Hanwen Shen, Asma Ghandeharioun, Craig Ferguson, Agata Lapedriza, Noah Jones, Shixiang Shane Gu, and Rosalind Picard. 2020. Human-centric dialog training via offline reinforcement learning. *arXiv preprint arXiv:2010.05848*.

Sooyeon Jeong, Sharifa Alghowinem, Laura Aymerich-Franch, Kika Arias, Agata Lapedriza, Rosalind Picard, Hae Won Park, and Cynthia Breazeal. 2020. A robotic positive psychology coach to improve college students' wellbeing. In *IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 187–194. IEEE.

Chuanyang Jin, Yutong Wu, Jing Cao, Jiannan Xiang, Yen-Ling Kuo, Zhiting Hu, Tomer Ullman, Antonio Torralba, Joshua B Tenenbaum, and Tianmin Shu. 2024. Mmtom-qa: Multimodal theory of mind question answering. *arXiv preprint arXiv:2401.08743*.

Margot E Kaminski, Matthew Rueben, William D Smart, and Cindy M Grimm. 2016. Averting robot eyes. *Md. L. Rev.*, 76:983.

Masahiro Kaneko and Danushka Bollegala. 2022. Unmasking the mask–evaluating social biases in masked language models. In *AAAI Conference on Artificial Intelligence*, volume 36, pages 11954–11962.

David A Kenny and Thomas Ledermann. 2010. Detecting, measuring, and testing dyadic patterns in the actor–partner interdependence model. *Journal of family psychology*, 24(3):359.

Muhammad Ali Khalidi. 2015. Three kinds of social kinds. *Philosophy and Phenomenological Research*, 90(1):96–112.

Hyunwoo Kim, Melanie Sclar, Xuhui Zhou, Ronan Le Bras, Gunhee Kim, Yejin Choi, and Maarten Sap. 2023. Fantom: A benchmark for stress-testing machine theory of mind in interactions. *arXiv preprint arXiv:2310.15421*.

Rodney Kinney, Chloe Anastasiades, Russell Authur, Iz Beltagy, Jonathan Bragg, Alexandra Buraczynski, Isabel Cachola, Stefan Candra, Yoganand Chandrasekhar, Arman Cohan, et al. 2023. The semantic scholar open data platform. *arXiv preprint arXiv:2301.10140*.

Max Kleiman-Weiner, Mark K Ho, Joseph L Austerweil, Michael L Littman, and Joshua B Tenenbaum. 2016. Coordinate to cooperate or compete: abstract goals and joint intentions in social interaction. In *CogSci*.

Dorit Kliemann, Ralph Adolphs, Tim Armstrong, Paola Galdi, David A Kahn, Tessa Rusch, A Zeynep Enkavi, Deuhua Liang, Steven Lograsso, Wenying Zhu, et al. 2022. Caltech conte center, a multimodal data resource for exploring social cognition and decision-making. *Scientific Data*, 9(1):138.

Chris Knight, Michael Studdert-Kennedy, and James Hurford. 2000. *The evolutionary emergence of language: social function and the origins of linguistic form*. Cambridge University Press.

Hubert Knoblauch and René Tuma. 2019. Videography and video analysis. *Sage research methods foundations. London: Sage*.

Vineet Kosaraju, Amir Sadeghian, Roberto Martín-Martín, Ian Reid, Hamid Rezatofighi, and Silvio Savarese. 2019. Social-bigat: Multimodal trajectory forecasting using bicycle-gan and graph attention networks. *Advances in Neural Information Processing Systems*, 32.

Jean Kossaifi, Robert Walecki, Yannis Panagakis, Jie Shen, Maximilian Schmitt, Fabien Ringeval, Jing Han, Vedhas Pandit, Antoine Toisoul, Björn Schuller, et al. 2019. Sewa db: A rich database for audio-visual emotion and sentiment research in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(3):1022–1040.

Grgur Kovač, Rémy Portelas, Peter Ford Dominey, and Pierre-Yves Oudeyer. 2023. The socialai school: Insights from developmental psychology towards artificial socio-cultural agents. *arXiv preprint arXiv:2307.07871*.

Jonathan Krause, Hailin Jin, Jianchao Yang, and Li Fei-Fei. 2015. Fine-grained recognition without part annotations. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5546–5555.

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.

Dennis Lee, Natasha Jaques, Chase Kew, Jiaxing Wu, Douglas Eck, Dale Schuurmans, and Aleksandra Faust. 2021. Joint attention for multi-agent coordination and social learning. *arXiv preprint arXiv:2104.07750*.

Nayeon Lee, Yejin Bang, Holy Lovenia, Samuel Cahyawijaya, Wenliang Dai, and Pascale Fung. 2023. Survey of social bias in vision-language models. *arXiv preprint arXiv:2309.14381*.

Sangmin Lee, Bolin Lai, Fiona Ryan, Bikram Boote, and James M. Rehg. 2024. Modeling multimodal social interactions: New challenges and baselines with densely aligned representations.

Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. Camel: Communicative agents for" mind" exploration of large language model society. *Advances in Neural Information Processing Systems*, 36:51991–52008.

Paul Pu Liang, Yun Cheng, Xiang Fan, Chun Kai Ling, Suzanne Nie, Richard Chen, Zihao Deng, Nicholas Allen, Randy Auerbach, Faisal Mahmood, et al. 2024. Quantifying & modeling multimodal interactions: An information decomposition framework. *Advances in Neural Information Processing Systems*, 36.

Paul Pu Liang, Chiyu Wu, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2021. Towards understanding and mitigating social biases in language models. In *International Conference on Machine Learning*, pages 6565–6576. PMLR.

Hugo Liu and Push Singh. 2004. Commonsense reasoning in and over natural language. In *Knowledge-Based Intelligent Information and Engineering Systems: 8th International Conference*, pages 293–306. Springer.

Weiwei Liu, Haobo Wang, Xiaobo Shen, and Ivor W Tsang. 2021. The emerging trends of multi-label learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):7955–7974.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Dan S Weld. 2019. S2orc: The semantic scholar open research corpus. *arXiv preprint arXiv:1911.02782*.

Sasha Luccioni, Christopher Akiki, Margaret Mitchell, and Yacine Jernite. 2024. Stable bias: Evaluating societal representations in diffusion models. *Advances in Neural Information Processing Systems*, 36.

Herbert G Lull. 1911. Moral instruction through social intelligence. *American Journal of Sociology*, 17(1):47–60.

Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. 2019. Dialoguernn: An attentive rnn for emotion detection in conversations. In *AAAI Conference on Artificial Intelligence*, volume 33, pages 6818–6825.

Arthur B Markman. 1989. Lms rules and the inverse base-rate effect: Comment on gluck and bower (1988).

Maja J Mataric. 1993. Designing emergent behaviors: From local interactions to collective intelligence.

Maja J Mataric. 1994. Learning to behave socially. *From animals to animats*, 3:453–462.

Maja J Matarić, Jon Eriksson, David J Feil-Seifer, and Carolee J Winstein. 2007. Socially assistive robotics for post-stroke rehabilitation. *Journal of Neuroengineering and Rehabilitation*, 4:1–9.

Leena Mathur, Ralph Adolphs, and Maja J Matarić. 2023a. Towards intercultural affect recognition: Audio-visual affect recognition in the wild across six cultures. In *2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–6. IEEE.

Leena Mathur, Maja Mataric, and Louis-Philippe Morency. 2023b. Expanding the role of affective phenomena in multimodal interaction research. In *Proceedings of the 25th International Conference on Multimodal Interaction*, pages 253–260.

Maya B Mathur and David B Reichling. 2016. Navigating a social world with robot partners: A quantitative cartography of the uncanny valley. *Cognition*, 146:22–32.

George J McCall. 2003. Interaction. *Handbook of Symbolic Interactionism*, 322:327–348.

Susan W McRoy and Graeme Hirst. 1993. Abductive explanation of dialogue misunderstandings. In *Sixth Conference of the European Chapter of the Association for Computational Linguistics*.

Louis-Philippe Morency. 2010. Modeling human communication dynamics [social sciences]. *IEEE Signal Processing Magazine*, 27(5):112–116.

Louis-Philippe Morency, Rada Mihalcea, and Payal Doshi. 2011. Towards multimodal sentiment analysis: Harvesting opinions from the web. In *International Conference on Multimodal Interfaces*, pages 169–176.

Louis-Philippe Morency, Ariadna Quattoni, and Trevor Darrell. 2007. Latent-dynamic discriminative models for continuous gesture recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE.

Masahiro Mori, Karl F MacDorman, and Norri Kageki. 2012. The uncanny valley [from the field]. *IEEE Robotics & automation magazine*, 19(2):98–100.

Katashi Nagao and Akikazu Takeuchi. 1994. Social interaction: Multimodal conversation with social agents. In *AAAI Conference on Artificial Intelligence*, volume 94, pages 22–28.

Clifford Nass and Youngme Moon. 2000. Machines and mindlessness: Social responses to computers. *Journal of Social Issues*, 56(1):81–103.

Rostislav Nedelchev, Jens Lehmann, and Ricardo Usbeck. 2020. Language model transformers as evaluators for open-domain dialogues. In *International Conference on Computational Linguistics*, pages 6797–6808.

David A Neequaye. 2023. Why rapport seems challenging to define and what to do about the challenge. *Collabra: Psychology*, 9(1).

Katherine Nelson. 2003. Self and social functions: Individual autobiographical memory and collective narrative. *Memory*, 11(2):125–136.

Evonne Ng, Hanbyul Joo, Liwen Hu, Hao Li, Trevor Darrell, Angjoo Kanazawa, and Shiry Ginosar. 2022. Learning to listen: Modeling non-deterministic dyadic facial motion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20395–20405.

Mihalis A Nicolaou, Vladimir Pavlovic, and Maja Pantic. 2014. Dynamic probabilistic cca for analysis of affective behavior and fusion of continuous annotations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1299–1311.

Jessie S Nixon, Sanne Poelstra, and Jacolien van Rij. 2022. Does error-driven learning occur in the absence of cues? examination of the effects of updating connection weights to absent cues. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 44.

Samuel A Olatunji, Vy Nguyen, Maya Cakmak, Aaron Edsinger, Charles C Kemp, Wendy A Rogers, and Harshal P Mahajan. 2024. Immersive participatory design of assistive robots to support older adults. *Ergonomics*, pages 1–15.

Mara Olekalns, Jeanne M Brett, and Laurie R Weingart. 2003. Phases, transitions and interruptions: Modeling processes in multi-party negotiations. *International Journal of Conflict Management*, 14(3/4):191–211.

Hae Won Park, Ishaan Grover, Samuel Spaulding, Louis Gomez, and Cynthia Breazeal. 2019. A model-free affective reinforcement learning approach to personalization of an autonomous social robot companion for early literacy education. In *AAAI Conference on Artificial Intelligence*, volume 33, pages 687–694.

Sungkyu Park, Marios Constantinides, Luca Maria Aiello, Daniele Quercia, and Paul Van Gent. 2020. Wellbeat: A framework for tracking daily well-being using smartwatches. *IEEE Internet Computing*, 24(5):10–17.

Soumitra Pathare, Alexandra Brazinova, and Itzhak Levav. 2018. Care gap: a comprehensive measure to quantify unmet needs in mental health. *Epidemiology and Psychiatric Sciences*, 27(5):463–467.

David Pautler and Alex Quilici. 1998. A computational model of social perlocutions. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 2*, pages 1020–1026.

Georgios Pavlakos, Ethan Weber, Matthew Tancik, and Angjoo Kanazawa. 2022. The one where they reconstructed 3d humans and environments in tv shows. In *European Conference on Computer Vision*, pages 732–749. Springer.

Catherine Pelachaud, Norman I Badler, and Mark Steedman. 1991. Linguistic issues in facial animation. In *Computer Animation'91*, pages 15–30. Springer.

Catherine Pelachaud, Norman I Badler, and Mark Steedman. 1996. Generating facial expressions for speech. *Cognitive science*, 20(1):1–46.

Stefano Pellegrini, Andreas Ess, and Luc Van Gool. 2010. Improving data association by joint modeling of pedestrian trajectories and groupings. In *European Conference on Computer Vision*, pages 452–465. Springer.

EL Pershina. 1986. Elementary contracts as a pragmatic basis of language interaction. In *Coling 1986 Volume 1: The 11th International Conference on Computational Linguistics*.

Martin J Pickering and Simon Garrod. 2021. *Understanding dialogue: Language use and social interaction*. Cambridge University Press.

Isabella Poggi and Francesca D'Errico. 2012. Social signals: a framework in terms of goals and beliefs. *Cognitive Processing*, 13:427–445.

Verónica Policarpo. 2015. What is a friend? an exploratory typology of the meanings of friendship. *Social Sciences*, 4(1):171–191.

Shrimai Prabhumoye, Rafal Kocielnik, Mohammad Shoeybi, Anima Anandkumar, and Bryan Catanzaro. 2021. Few-shot instruction prompts for pretrained language models to detect social biases. *arXiv preprint arXiv:2112.07868*.

Xavier Puig, Eric Undersander, Andrew Szot, Mikael Dallaire Cote, Tsung-Yen Yang, Ruslan Partsey, Ruta Desai, Alexander William Clegg, Michal Hlavac, So Yeon Min, et al. 2023. Habitat 3.0: A co-habitat for humans, avatars and robots. *arXiv preprint arXiv:2310.13724*.

Jeffrey J Rachlinski. 1998. The limits of social norms. *Chi.-Kent L. Rev.*, 74:1537.

Stephen J Read, Chadwick J Snow, and Dan Simon. 2013. Constraint satisfaction processes in social reasoning. In *Proceedings of the 25th Annual Cognitive Science Society*, pages 964–969. Psychology Press.

Donald Redfoot, Lynn Feinberg, and Ari N Houser. 2013. *The aging of the baby boom and the growing care gap: A look at future declines in the availability of family caregivers*, volume 12. AARP Public Policy Institute Washington, DC.

Frieda Rong and Max Kleiman-Weiner. 2024. Value internalization: Learning and generalizing from social reward. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 46.

David Rozenberszki, Or Litany, and Angela Dai. 2022. Language-grounded indoor 3d semantic segmentation in the wild. In *European Conference on Computer Vision*, pages 125–141. Springer.

Ognjen Rudovic, Jaeryoung Lee, Miles Dai, Björn Schuller, and Rosalind W Picard. 2018. Personalized machine learning for robot perception of affect and engagement in autism therapy. *Science Robotics*, 3(19):eaao6760.

Aneesha Sampath, Victoria Lin, and Louis-Philippe Morency. 2022. Seedbert: Recovering annotator rating distributions from an aggregated label. *arXiv preprint arXiv:2211.13196*.

Akane Sano. 2016. *Measuring college students' sleep, stress, mental health and wellbeing with wearable sensors and mobile phones*. Ph.D. thesis, Massachusetts Institute of Technology.

Sebastin Santy, Jenny T Liang, Ronan Le Bras, Katharina Reinecke, and Maarten Sap. 2023. Nlpositionality: Characterizing design biases of datasets and models. *arXiv preprint arXiv:2306.01943*.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A Smith, and Yejin Choi. 2019a. Social bias frames: Reasoning about social and power implications of language. *arXiv preprint arXiv:1911.03891*.

Maarten Sap, Ronan LeBras, Daniel Fried, and Yejin Choi. 2022. Neural theory-of-mind? on the limits of social intelligence in large lms. *arXiv preprint arXiv:2210.13312*.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. 2019b. Socialiqa: Commonsense reasoning about social interactions. *arXiv preprint arXiv:1904.09728*.

Allison Sauppé and Bilge Mutlu. 2015. The social impact of a robot co-worker in industrial settings. In *Proceedings of the 33rd annual ACM Conference on Human Factors in Computing Systems*, pages 3613–3622.

Brian Scassellati, Henny Admoni, and Maja Matarić. 2012. Robots for use in autism research. *Annual Review of Biomedical Engineering*, 14:275–294.

Björn Schuller, Michel Valster, Florian Eyben, Roddy Cowie, and Maja Pantic. 2012. Avec 2012: the continuous audio/visual emotion challenge. In *ACM International Conference on Multimodal Interaction*, pages 449–456.

John Searle. 2010. *Making the social world: The structure of human civilization*. Oxford University Press.

John R Searle. 1995. *The construction of social reality*. Simon and Schuster.

John R Searle. 1998. Social ontology and the philosophy of society. *Analyse & Kritik*, 20(2):143–158.

John R Searle. 2012. Human social reality and language. *Phenomenology and Mind*, (2):24–33.

Natalie Shapira, Mosh Levy, Seyed Hossein Alavi, Xuhui Zhou, Yejin Choi, Yoav Goldberg, Maarten Sap, and Vered Shwartz. 2023. Clever hans or neural theory of mind? stress testing social reasoning in large language models. *arXiv preprint arXiv:2305.14763*.

Ashish Sharma, Inna W Lin, Adam S Miner, David C Atkins, and Tim Althoff. 2023. Human–ai collaboration enables more empathic conversations in text-based peer-to-peer mental health support. *Nature Machine Intelligence*, 5(1):46–57.

Erez Shmueli, Vivek K Singh, Bruno Lepri, and Alex Pentland. 2014. Sensing, understanding, and shaping social behavior. *IEEE Transactions on Computational Social Systems*, 1(1):22–34.

Michael Shum, Max Kleiman-Weiner, Michael L Littman, and Joshua B Tenenbaum. 2019. Theory of minds: Understanding behavior in groups through inverse planning. In *AAAI Conference on Artificial Intelligence*, volume 33, pages 6163–6170.

Aaditya K Singh and DJ Strouse. 2024. Tokenization counts: the impact of tokenization on arithmetic in frontier llms. *arXiv preprint arXiv:2402.14903*.

John J Skowronski, Andrew L Betz, Charles P Thompson, and Laura Shannon. 1991. Social memory in everyday life: Recall of self-events and other-events. *Journal of Personality and Social Psychology*, 60(6):831.

Mark Snyder and Nancy Cantor. 1980. Thinking about ourselves and others: Self-monitoring and social knowledge. *Journal of Personality and Social Psychology*, 39(2):222.

Mohammad Soleymani, Kalin Stefanov, Sin-Hwa Kang, Jan Ondras, and Jonathan Gratch. 2019. Multimodal analysis and estimation of intimate self-disclosure. In *ACM International Conference on Multimodal Interaction*, pages 59–68.

Samuel Lee Spaulding. 2022. *Lifelong Personalization for Social Robot Learning Companions: Interactive Student Modeling Across Tasks and Over Time*. Ph.D. thesis, Massachusetts Institute of Technology.

Julia G Stapels, Angelika Penner, Niels Diekmann, and Friederike Eyssel. 2023. Never trust anything that can think for itself, if you can't control its privacy settings: The influence of a robot's privacy settings on users' attitudes and willingness to self-disclose. *International Journal of Social Robotics*, 15(9):1487–1505.

Kim Sterelny. 2007. Social intelligence, human intelligence and niche construction. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 362(1480):719–730.

Kyle Strabala, Min Kyung Lee, Anca Dragan, Jodi Forlizzi, and Siddhartha S Srinivasa. 2012. Learning the communication of intent prior to physical collaboration. In *The 21st IEEE International Symposium on Robot and Human Interactive Communication*, pages 968–973. IEEE.

Ruth Strang. 1930. Measures of social intelligence. *American Journal of Sociology*, 36(2):263–269.

William R Swartout, Jonathan Gratch, Randall W Hill Jr, Eduard Hovy, Stacy Marsella, Jeff Rickel, and David Traum. 2006. Toward virtual humans. *AI Magazine*, 27(2):96–96.

Leila Takayama, Wendy Ju, and Clifford Nass. 2008. Beyond dirty, dangerous and dull: what everyday people think robots should do. In *ACM/IEEE International Conference on Human-Robot Interaction*, pages 25–32.

Diana I Tamir and Brent L Hughes. 2018. Social rewards: from basic social building blocks to complex social behavior. *Perspectives on Psychological Science*, 13(6):700–717.

Ada V Taylor, Ellie Mamantov, and Henny Admoni. 2022. Observer-aware legibility for social navigation. In *IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 1115–1122. IEEE.

Edward L Thorndike. 1920. Intelligence and its uses. *Harper's Magazine*, 140:227–235.

Robert L Thorndike and Saul Stein. 1937. An evaluation of the attempts to measure social intelligence. *Psychological Bulletin*, 34(5):275.

Linda Tickle-Degnen and Robert Rosenthal. 1990. The nature of rapport and its nonverbal correlates. *Psychological inquiry*, 1(4):285–293.

Sabine Trepte. 2013. Social identity theory. In *Psychology of entertainment*, pages 255–271. Routledge.

Jonathan H Turner. 1988. *A theory of social interaction*. Stanford University Press.

Jay J Van Bavel, JY Xiao, and Leor M Hackel. 2013. Social identity shapes social perception and evaluation. *Neuroscience of prejudice and intergroup relations*, 110.

Linda J Van Hamme and Edward A Wasserman. 1994. Cue competition in causality judgments: The role of nonpresentation of compound stimulus elements. *Learning and motivation*, 25(2):127–151.

Alessandro Vinciarelli and Anna Esposito. 2018. Multimodal analysis of social signals. In *The Handbook of Multimodal-Multisensor Interfaces: Signal Processing, Architectures, and Detection of Emotion and Cognition-Volume 2*, pages 203–226.

Alessandro Vinciarelli, Maja Pantic, and Hervé Bourlard. 2009. Social signal processing: Survey of an emerging domain. *Image and Vision Computing*, 27(12):1743–1759.

Alessandro Vinciarelli, Maja Pantic, Dirk Heylen, Catherine Pelachaud, Isabella Poggi, Francesca D'Errico, and Marc Schroeder. 2011. Bridging the gap between social animal and unsocial machine: A survey of social signal processing. *IEEE Transactions on Affective Computing*, 3(1):69–87.

Joshua Wainer, David J Feil-Seifer, Dylan A Shell, and Maja J Mataric. 2006. The role of physical embodiment in human-robot interaction. In *IEEE International Symposium on Robot and Human Interactive Communication*, pages 117–122. IEEE.

Hui Wang, Jinying Lin, Zhen Ma, Yurii Vasylkiv, Heike Brock, Keisuke Nakamura, Randy Gomez, Bo He, and Guangliang Li. 2022. Affective behavior learning for social robot haru with implicit evaluative feedback. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3881–3888. IEEE.

Zhen Wang, Zhao Song, Chen Shen, and Shuyue Hu. 2023. Emergence of punishment in social dilemma with environmental feedback. In *AAAI Conference on Artificial Intelligence*, volume 37, pages 11708–11716.

Paul Watzlawick, Janet Beavin Bavelas, and Don D Jackson. 2011. *Pragmatics of human communication: A study of interactional patterns, pathologies and paradoxes*. WW Norton & Company.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Susanne Weis. 2008. *Theory and Measurement of Social Intelligence as a Cognitive Performance Construct*. Ph.D. thesis.

Susanne Weis and Heinz-Martin Süß. 2005. Social intelligence–a review and critical discussion of measurement concepts.

Alex Wilf, Leena Mathur, Sheryl Mathew, Claire Ko, Youssouf Kebe, Paul Pu Liang, and Louis-Philippe Morency. 2023. Social-iq 2.0 challenge: Benchmarking multimodal social understanding.

Edward O Wilson. 2012. *The social conquest of earth*. WW Norton & Company.

Marc Wittmann. 2011. Moments in time. *Frontiers in Integrative Neuroscience*, 5:66.

Sarah A Wu, Rose E Wang, James A Evans, Josh Tenenbaum, David C Parkes, and Max Kleiman-Weiner. 2020. Too many cooks: Coordinating multi-agent collaboration through inverse planning. In *CogSci*.

Georgios N Yannakakis, Roddy Cowie, and Carlos Busso. 2017. The ordinal nature of emotions. In *International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 248–255. IEEE.

Georgios N Yannakakis and Héctor P Martínez. 2015. Ratings are overrated! *Frontiers in ICT*, 2:13.

Amir Zadeh, Michael Chan, Paul Pu Liang, Edmund Tong, and Louis-Philippe Morency. 2019. Social-iq: A question answering benchmark for artificial social intelligence. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8807–8817.

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2236–2246.

Leslie A Zebrowitz. 1990. *Social perception.* Thomson Brooks/Cole Publishing Co.

Andy Zeng, Maria Attarian, Brian Ichter, Krzysztof Choromanski, Adrian Wong, Stefan Welker, Federico Tombari, Aveek Purohit, Michael Ryoo, Vikas Sindhwani, et al. 2022. Socratic models: Composing zero-shot multimodal reasoning with language. *arXiv preprint arXiv:2204.00598*.

Yuanhao Zhai, Mingzhen Huang, Tianyu Luan, Lu Dong, Ifeoma Nwogu, Siwei Lyu, David Doermann, and Junsong Yuan. 2023. Language-guided human motion synthesis with atomic actions. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 5262–5271.

Angie Zhang, Olympia Walker, Kaci Nguyen, Jiajun Dai, Anqing Chen, and Min Kyung Lee. 2023a. Deliberating with ai: Improving decision-making for the future through participatory ai design and stakeholder deliberation. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1):1–32.

Hongxin Zhang, Weihua Du, Jiaming Shan, Qinhong Zhou, Yilun Du, Joshua B Tenenbaum, Tianmin Shu, and Chuang Gan. 2023b. Building cooperative embodied agents modularly with large language models. *arXiv preprint arXiv:2307.02485*.

Yaqi Zhang, Di Huang, Bin Liu, Shixiang Tang, Yan Lu, Lu Chen, Lei Bai, Qi Chu, Nenghai Yu, and Wanli Ouyang. 2023c. Motiongpt: Finetuned llms are general-purpose motion generators. *arXiv preprint arXiv:2306.10900*.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. 2024. Memorybank: Enhancing large language models with long-term memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19724–19731.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, et al. 2023. Sotopia: Interactive evaluation for social intelligence in language agents. *arXiv preprint arXiv:2310.11667*.

Caleb Ziems, Jane Dwivedi-Yu, Yi-Chia Wang, Alon Halevy, and Diyi Yang. 2023. NormBank: A knowledge bank of situational social norms. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7756–7776, Toronto, Canada. Association for Computational Linguistics.

A Search Queries and Filtering Criteria

We accessed data from Semantic Scholar (Lo et al., 2019; Kinney et al., 2023), a scientific research tool from the Allen Institute for AI. We bulk-downloaded meta-data of papers that (1) had "Computer Science" as one of its fields and (2) satisfied the following Semantic Scholar query: `["social" AND "model"] OR ["social" and "agent"] OR ["social" and "artificial intelligence"] OR ["social" and "machine learning"]`. This initial search yielded approximately 161,000 papers. We, then, filtered our sample to include relevant papers from venues in the "top 20" per field by Google Scholar's venue-listings across the 6 computing communities of artificial intelligence (this Google Scholar listing includes key machine learning conferences), computer vision, natural language processing, robotics, and human-computer interaction (this Google Scholar listing includes human-machine interaction and human-robot interaction). This filtering step yielded 3,257 papers.

On the assumption that the papers between 2010 and 2020 would be the most influential in ascertaining priorities for each field and papers from 2020 onwards capture the most recent priorities of each field, we sampled 20 papers per computing community. 40% of papers were from 2010-2019, and 60% of papers were from 2020-2023 (15% per year). Retaining for relevance, we pick the most cited 8 papers from the past decade and the most cited 3 papers from each year of 2020-2023. We examined these 120 papers to help ensure that our discussion of prior work captures research priorities of computing communities in Social-AI.

B Social-AI Infrastructure

We discuss several data infrastructure considerations to supplement Social-AI research efforts towards the core technical challenges in our paper.

Social-AI Data Sources Commonly-used sources of data for Social-AI model training and evaluation are *static* (e.g., lab-collected interactions, TV shows) and do not contain a naturalistic distribution of long-tail social phenomena, nuanced social signals, multi-perspective dynamics, and other dimensions of social context visualized in Figure 1B. The majority of online video data, for example, are sociotechnical constructions (Knoblauch and Tuma, 2019), influenced by choices such as frame angles and scene cuts.

Social-AI research may benefit from emphasizing richer sources of naturalistic interactions, as well as *interactive* sources of data. Examples of recent work exploring richer sources of social interaction data include multimodal egocentric perception (Grauman et al., 2023) across audio, motion/position, gaze, vision, pose, and natural language, as well as a multi-year study of multimodal neural and psychological factors related to human social cognition (Kliemann et al., 2022). Multimodal, ethically-collected data sources that pair natural language, vision, and other neural and sensory modalities (e.g., olfaction, physiology) during social interactions can be invaluable sources of data to advance Social-AI research. We also believe there is great potential to explore techniques for re-purposing and augmenting existing social interaction datasets to supplement the curation of real-world datasets (e.g., reconstructing 3D humans from 2D videos) (Pavlakos et al., 2022).

Social-AI Benchmarks Curating multimodal datasets of *interactions in-the-wild* that include Social-AI agents and actors situated within multiple dimensions of social context will be useful to advancing Social-AI research. There is a scarcity of shared, longitudinal datasets of real-world interactions across social contexts that can support researchers in stress-testing algorithms to address core technical challenges in Social-AI research. For example, a researcher developing an algorithm to address the challenge of modeling multiple perspectives during social interactions might find it useful to explore how effective the approach is when tested for the same actor in different *interaction structures*, spanning dyadic, multi-party, and group contexts. There is currently no shared infrastructure for systematically performing these experiments to inform our understanding of Social-AI agent performance along various dimensions of social context. This type of understanding is important to safely deploy Social-AI agents in-the-wild.

Social-AI Data Annotation Obtaining fine-grained annotations of multiple social constructs during social interactions is a resource-intensive task. In the case of social interaction video data, a fine-grained annotation might include labels at different hierarchical social units (e.g., behaviors, events) that may overlap in video segments. Each minute of an interaction can take more than an hour for a trained annotator to analyze and annotate (Morency, 2010). It is important to include multiple annotators for each sample and fairly compensate annotators for these types of cognitively-demanding annotations (Huang et al., 2023).

The challenge of ambiguity in social constructs amplifies the difficulty of this annotation process, since annotators can interpret constructs differently, even when provided with the same initial annotation instructions. It is important to design annotation paradigms that consciously include annotators from diverse backgrounds, in order to represent multiple perspectives on social constructs that are likely to exist across gender, age, culture, nationality, and other demographic factors (Ding et al., 2022; Santy et al., 2023).

In addition, annotators' social perception speeds can also differ when interpreting social interactions, causing *temporal delays* between occurrences of social phenomena and labels across annotators. Techniques to address these temporal delays in annotations must be developed and standardized in modeling paradigms by Social-AI researchers.

Community Resource To accompany the paper and stimulate community research, we contribute a repository of resources for researchers interested in addressing Social-AI challenges: `https://github.com/l-mathur/social-ai`. This repository contains links to papers, books, doctoral dissertations, datasets, benchmarks, simulation environments, and university courses relevant to Social-AI research. This resource will be continually updated for the research community.