Towards Real-World Social AI

Paul Pu Liang
Carnegie Mellon University

1 Introduction

As intelligent systems increasingly blend into our everyday lives, building AI systems that display social intelligence has become one of the next grand challenges in the field. **Socially intelligent AI** should comprehend human social cues, intents, and affective states, engage in social conversation, and understand social norms and commonsense in order to maintain a rich level of interpersonal interaction with humans. Social intelligence is currently a defining trait uniquely natural to humans. In everyday social interactions, we convey our intentions through a coordinated structure of multimodal signals: language (words, phrases, and sentences), visual (gestures and expressions), and acoustic (paralinguistics and changes in vocal tones). While AI has shown tremendous promise in solving various tasks, designing social AI with the capability to communicate the same way humans do, by incorporating all involved modalities, is a fundamental research challenge.

My research builds towards **real-world social AI** that understands and engages in human communication, thereby narrowing the gap in computers' understanding of humans and opening new horizons for the creation of socially intelligent entities. The creation of social AI would bring about real-world advances in human sensing and robot design, with the end goal of engaging people through social and physical interactions [14], monitoring human behavior to understand and predict the types of help people need [8, 10], and offering assistance in schools, hospitals, and the workplace [12]. While prior research towards social intelligence has made impressive strides in affective computing and dialog systems, my research further focuses on bridging the gap towards real-world deployment. I strongly believe that real-world social AI has the capability to democratize access to important human-centric areas such as healthcare and education [12, 16]. Therefore, my vision focuses on ensuring the accessibility of such models so that no social group will be at a disadvantage when deployed, particularly underrepresented groups [17]. As steps towards real-world social AI, I have outlined three major milestones:

1. Multimodal perception of human communication: Human communication is a structured system of social signals used by humans to convey their intentions using both verbal and nonverbal channels. This constitutes a significant portion of human social behavior, including face-to-face conversation, video chatting, and multimedia opinion sharing. The first step aims to **perceive multimodal communication** from a human speaker by comprehending social cues, intents, affective states, personalities, and references to the broader environment.

2. Interactive social intelligence: Step two involves modeling the interactive loop between social perception and action - both **communicating** with humans through multimodal behaviors and **acting** in an embodied environment. This interactive loop between listening and responding to social behaviors happens over a long-term horizon.

3. Towards real-world social intelligence: Closing the gap towards real-world deployment via 1) **robust** learning in the face of noisy and missing modalities, 2) **fair** representation learning from human-centric data, and 3) **interpretable** modeling of social commonsense.

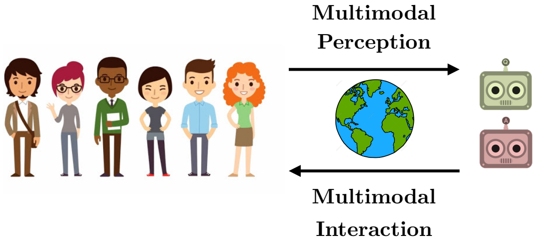

Figure 1: My research towards real-world social AI involves 1) multimodal perception of human communication, 2) modeling the interactive loop between multimodal perception and action, and 3) ensuring robustness, fairness, and interpretability in real-world applications.

My Ph.D. research aims to take a major step towards real-world social AI. I am tremendously dedicated to my research vision and have made significant contributions towards multimodal perception and robustness. This fellowship will enable me to push my research vision further towards more interactive and fair social AI. From a broader perspective, the outcome of my research will also present fundamental theoretical and practical insights in multimodal learning, allowing researchers to design models that capture the benefits of multimodal data sources and deploy them in the real world. In the following sections, I describe my existing research and outline directions for future work.

2 Multimodal Perception of Human Communication

From a computational perspective, the modeling of human communication across both verbal and nonverbal behaviors focuses on tasks such as multimodal sentiment analysis [18], emotion recognition [3], and personality traits recognition [20]. To comprehend human communication, there is a need for 1) large multimodal resources with diversity in training samples, topics, speakers, and annotations, as well as 2) powerful models for multimodal communication. I have made significant open-source contributions in these areas.

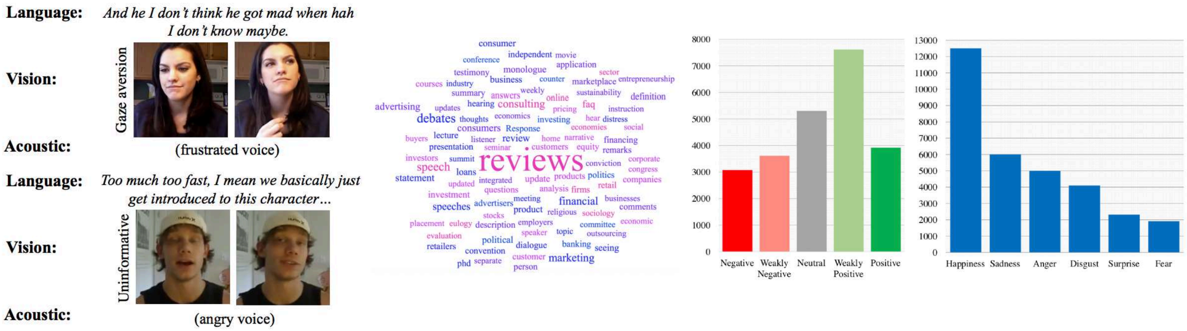

Figure 2: I have worked towards addressing the lack of multimodal resources by collecting and releasing the largest dataset of multimodal sentiment and emotion recognition with diversity in training samples, topics, speakers, and annotations, thereby enabling generalizable studies of human communication.

2.1 Multimodal Communication Resources

2.1.1 Preliminary work in multimodal resources: I have worked towards addressing the lack of multimodal resources by **collecting and releasing the largest dataset of multimodal sentiment and emotion recognition enabling generalizable studies of human communication** (see Figure 2). CMU-MOSEI contains 23,500 annotated video segments from 1,000 distinct speakers and 250 topics. The diversity in topics, speakers, annotations, and modalities allows for generalizable studies of speaker and topic-independent features. The multimodal dataset and a general multimodal data loading framework are provided to the scientific community to encourage valuable research in human communication analysis. This work culminated in an oral presentation at ACL 2018 [24] and was also the bulk of my master's thesis at CMU [9]. Since then, the dataset has also been the subject of two workshop challenges in modeling human multimodal language at ACL 2018 and ACL 2020, and has been a standard benchmark dataset for the multimodal machine learning community.

2.2 Multimodal Communication Modeling

My research has substantially improved the state-of-the-art in human communication modeling. These approaches have also contributed to core technical algorithms for multimodal learning from heterogeneous data and have been used by the broader research community for applications in conversational agents [19], depression detection [1], face alignment [6], personalized recommendation [23], and medical image segmentation [5].

2.2.1 Preliminary work in multimodal representation learning: My research has contributed to **better understandings of the desiderata for multimodal representations beyond discriminative performance** [22]. I proposed an approach based on factorizing multimodal representations into multimodal discriminative and modality-specific generative factors. The discriminative factor learns joint features across modalities that achieve state-of-the-art performance for affect analysis. At the same time, the modality-specific generative factors enable controllable generation of human language based on individual factors, better model partially missing modalities, and allow analysis of local contributions from each modality during prediction, desiderata previously unachievable via purely discriminative approaches.

2.2.2 Preliminary work in multimodal modeling: I have made **contributions towards data and compute-efficient multimodal learning via simple yet strong models** [11]. These models are based on stronger statistical baselines rather than black-box neural networks. Our approach assumes a fully-factorized probabilistic generative model of multimodal data from a latent representation. Careful model design allows us to capture expressive unimodal, bimodal, and trimodal interactions while at the same time retaining simplicity and efficiency during learning and inference. These models show strong performance on both supervised and semi-supervised multimodal prediction, as well as significant (10 times) speedups over neural models during inference.

2.2.3 Preliminary work in multimodal affect analysis: I proposed a **new perspective on human-centric affect analysis by modeling both person-independent and person-dependent signals** [10]. Some emotional expressions are almost universal person-independent behaviors and can be recognized directly from a video. For example, an open mouth with raised eyebrows and a loud voice is likely to be associated with surprise. However, emotions are also expressed in a person-dependent fashion with idiosyncratic behaviors where it may not be possible to directly estimate absolute emotion intensities. Instead, it would be easier to compare two video segments of the same person and judge whether there was a relative change in emotion intensities. For example, a person could have naturally furrowed eyebrows and we should not always interpret this as a display of anger, but rather compare two video segments to determine relative changes in anger. I designed a model combining both signals to achieve state-of-the-art audio-visual emotion recognition performance and allow for fine-grained investigation of person-independent and person-dependent behaviors.

3 Interactive Social Intelligence

One of my core research thrusts for the remaining duration of my PhD lies in modeling the **long-term interactive loop between social perception and action**. After perceiving human communication, social AI agents should have the ability to **communicate** through multimodal behaviors and **act** in an embodied environment. It remains a core technical challenge to model the high-dimensional action spaces as well as generation quality of multimodal outputs.

3.0.1 Ongoing work in multimodal reinforcement learning: I am currently working on **action abstractions in multimodal reinforcement learning which will enable agents to communicate and act over long-term horizons**. These environments have large multimodal state and action spaces consisting of both actions as well as text from the environment or dialog with other agents. This makes them especially challenging for existing RL algorithms since it is intractable to enumerate over large action spaces. Action abstractions allow us to learn low-dimensional, semantically meaningful representations which can then be decoded into raw high-dimensional actions. I aim to scale up action abstractions for challenging discrete and continuous environments by appropriate design and optimization of a good latent action space.

Stage	Agents	Data collection	Learning algorithm	Evaluation metric
1. Imitation	AI	Human demonstration	Supervised learning	Generation likelihood
2. Self-play	AI with AI	Simulated environment	Reinforcement learning	Cumulative reward & Generation likelihood
3. Interaction	AI with Human	Simulated environment & Human-in-the-loop labeling	Reinforcement learning & Active learning	Human judgement & Generation likelihood

Table 1: I envision a new standard of evaluation benchmarks that increasingly assess the realism of interactive social AI across imitation, self-play, and interaction stages, each building on top of the previous stage. I also plan to collect more realistic interactive benchmarks that better represent real-world social AI in dialogue, robotics, and healthcare.

3.0.2 Long-term goals in benchmarking interactive intelligence: Any AI model can only be as good as the metrics used to evaluate it. Therefore, I plan to **collect realistic interactive benchmarks to better evaluate real-world social AI**. Existing benchmarks lack the component of interactivity and focus primarily on supervised tasks. Instead, I envision a new series of evaluation **microtasks** that each specialize in a subset of social interactions. Each of these microtasks increasingly assess the realism of social intelligence, categorized according to the agents involved, data collection process, learning algorithms, and evaluation metrics as illustrated in Table 1. Stage 1, the **imitation** stage, tests whether AI is able to imitate humans in social settings. Most of the existing work in supervised affect recognition and dialog modeling falls under this category. Stage 2, the **self-play** stage, tests whether AI is able to interactively engage with itself. A core technical challenge in this stage lies in tackling the problem of language drift where core knowledge of human language is forgotten and both agents descend into language model regions of low likelihood [7]. Stage 3, the **interactive** stage, tests whether AI is able to engage in interactive social communication with a real human. I aim to leverage human-in-the-loop learning and active learning to provide useful human labels in an interactive multimodal setting. The culmination of these 3 stages will more accurately benchmark the interactive capabilities of social AI and uncover the shortcomings of existing models.

To realize this long-term goal, I plan to first focus on a specific aspect of social interaction: multimodal dialog between a human speaker and actor in a situated environment. Multimodal dialog is a good testbed due to the use of nonverbal gestures in addition to language as well as references to the broader environment. However, current multimodal dialog datasets are limited by size, modalities (it has been shown that models rely primarily on language), and actions (participants primarily talk without performing actions in an environment). To mitigate these shortcomings, I plan to explore data-collection methods that involve scripted situations with actors that engage nonverbal modalities beyond language and act in the environment. Furthermore, I will also focus on better evaluation metrics for actions beyond those recorded in the dataset.

4 Towards Real-world Social Intelligence

To enable social AI technologies for real-world deployment, I have identified 3 core challenges that must be adequately addressed: 1) **robustness to noisy and missing modalities**, 2) **fair representation learning from human-centric data** and 3) **interpretable modeling of social commonsense**. I am also currently working towards real-world applications of social AI in healthcare by tacking problems such as privacy-preserving learning from mobile data.

4.1 Robustness to Noisy and Missing Modalities

While social AI requires modeling of human communication, real-world multimodal data is often imperfect as a result of missing entries, noise corruption, or missing modalities entirely. Human-centric data is also often imperfect due to personal idiosyncrasies which affect the contribution of certain modalities during social interactions. For example, multimodal dialogue systems trained on acted TV shows are susceptible to poor performance when deployed in the real world where users might be less expressive in using facial gestures. This calls for robust models that can still make accurate predictions despite only having access to a (possibly noisy) subset of signals.

4.1.1 Preliminary work in handling noisy modalities: I proposed a **mathematically grounded tensor representation learning method to deal with noisy modalities** in time-series data (e.g. text, videos, audio) [13]. This method is based on the observation that multimodal time series data often exhibits correlations across time and modalities which lead to low-rank multimodal representations. However, the presence of noise or incomplete values breaks these correlations and results in tensor representations of higher rank. Regularizing the rank of tensor representations therefore provides a denoising effect and our model achieves strong results across various levels of imperfection.

4.1.2 Preliminary work in handling missing modalities: I also investigated the scenario where entire modalities may be missing during deployment [21]. Existing methods always learn a joint representation with all modalities as input, making them susceptible when modalities are not all available. My method is based on the insight that translation from a source to target modality learns joint representations using only the source modality as input while extracting information present in the target modality. This **new paradigm of multimodal learning by translating between modalities requires only the source modality at test time which ensures robustness to target modalities**. Experiments show robust performance in multimodal sentiment and emotion analysis while requiring only language as the source, often comparable to models operating on all 3 modalities (language, visual, and audio).

4.1.3 Long-term goals in achieving dynamic robustness: A challenge not addressed by previous work is **scenarios where modalities are dynamically missing and noisy over long-term social interactions**. In this setting, the imperfect modalities are unknown and dynamically change at different times of model usage, which better represents real-world multimodal learning where reliable data sources constantly change (see Figure 3). I plan to first collect realistic interactive datasets that better represent real-world social AI with imperfect modalities for multimodal dialogue, human-robot interaction, and healthcare diagnosis. Tackling dynamic robustness over long signals also carries several technical challenges involving temporal credit assignment, assessing the utility of each modality, and formalizing the tradeoffs between unimodal and multimodal learning in terms of performance and robustness metrics (see section 4.2.3).

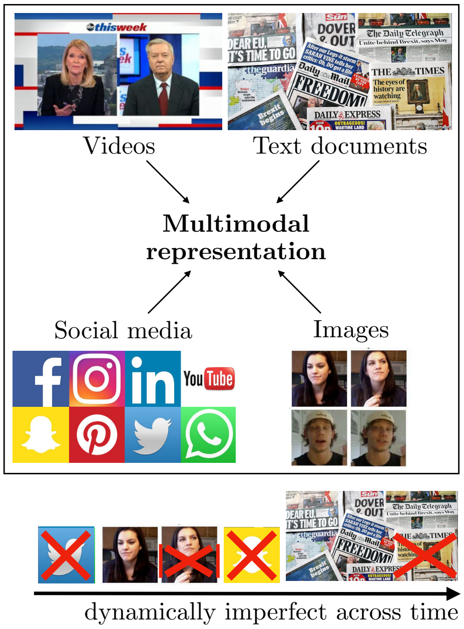

Figure 3: Real-world multimodal learning suffers when the imperfect modalities are unknown and dynamically change at different times of model usage. I plan to collect realistic benchmarks with imperfect modalities in dialogue, robotics, and healthcare applications.

4.2 Fair Human-centric Multimodal Learning

To safely deploy human-centric social AI models in real-world scenarios such as healthcare, legal systems, and social science, it is necessary to recognize the role they play in shaping social biases and stereotypes. Previous work has revealed the presence of undesirable biases in word embeddings involving gender, race, and religion. Similarly, we must carefully characterize these biases and design algorithms to mitigate biases for multimodal social AI models.

4.2.1 Preliminary work in debiasing sentence representations: While existing methods for debiasing word embeddings are largely successful [2], there has been a shift from word embeddings to contextual sentence representations such as ELMo and BERT. We are **the first to investigate the presence of social biases in these sentence-level representations and propose a new method, Sent-Debias, to mitigate these biases** [15]. Representational biases are harmful biases resulting from stereotyping that propagates negative generalizations about particular social groups. These are currently measured using a set of word association tests between predefined social constructs (e.g. gender and racial terms) and social professions (e.g. occupations, academic fields). Sent-Debias is based on contextualizing bias-attribute words (e.g. man, woman) using a diverse set of sentence templates into bias-attribute sentences. Our experiments showed the importance of 1) naturally-occurring sentence templates from large text corpora over simple templates, and 2) both the quality and quantity of sentence templates used. Our method reduces bias effect size while retaining performance on sentiment analysis, linguistic acceptability, and language understanding.

4.2.2 Ongoing work in debiasing general modalities: To **quantify and mitigate biases in social AI systems that learn from multimodal experience**, we must further contextualize bias-attribute words into their corresponding entities in the target modalities by modeling the highly complex relationships between modalities (see Figure 4). Furthermore, contextualization could introduce potential biases as well: standard image retrieval models do discriminate against women and people of color (e.g. given the word "scientist", retrieving a higher proportion of male than female images). I am working towards these directions to effectively mitigate bias in modern AI technologies such as pretrained language models, cross-modal retrieval models, and human-centric prediction models.

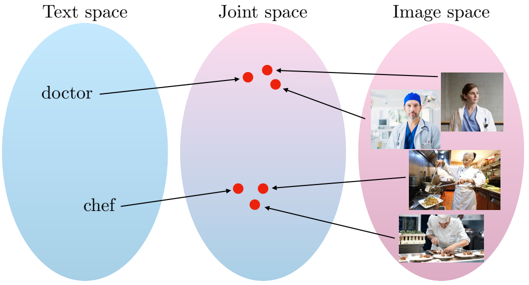

Figure 4: Learning a joint space across multimodal data allows us to accurately contextualize words into their corresponding entities in the target modalities for bias estimation and removal. It is also important to mitigate potential biases during contextualization.

4.2.3 Long-term goals in formalizing tradeoffs in multimodal learning: Existing approaches primarily optimize for prediction performance from multimodal data sources without formally quantifying the tradeoffs between improved performance and the potential drawbacks involving increased time and space complexity of learning from another modality, risk of decreased robustness from imperfect modalities, and risk of unfair learning from biased modalities. For example, training supervised models to predict human affect from multimodal data can lead to an over-reliance on the most informative language modality, which makes these models highly sensitive to language imperfections and social biases in language embeddings during testing. On the theoretical side, my goal is to **formally characterize the desiderata in multimodal representation learning balancing various performance, complexity, robustness, and fairness metrics**. At the same time, I also plan to empirically quantify these trade-offs to determine the overall contribution of a modality given the potential drawbacks on a set of real-world multimodal benchmarks spanning healthcare, robotics, and affective computing. Answering these questions will bridge the gap towards real-world social AI that captures the benefits of multimodal data sources while accurately considering and mitigating the potential risks involving robustness and fairness.

4.3 Interpretable Modeling of Social Commonsense

4.3.1 Long-term goals in building social commonsense: Social commonsense reflects a collective measure of self and social awareness, evolved social beliefs and attitudes, and a general set of social norms governing everyday interactions. It is challenging for AI to learn social commonsense directly from data since such basic knowledge is often not reflected in our pursuit of increasingly complex real-world datasets. I aim to **impart social commonsense into AI agents by designing interpretable structures of social knowledge as priors for social AI**. The language of knowledge graphs allows for interpretable modeling of social commonsense via a set of entities representing social behaviors and relations representing social interactions. As an example, entities could represent individual behaviors such as the pitch of voice or the presence of a smile, while relations would represent either actions from the same agent (smile followed by a nod) or actions from one agent to another (saying a phrase and obtaining a response). Humans, via theory of mind, have strong psychological priors of social commonsense and can accurately predict the result of engaging via one form of interaction given a currently observed set of social behaviors. Similarly, in building social AI, inference under a subset of entities and relations in such a knowledge graph should result in a deduction of the social states resulting from executing those interactions given those behaviors. An interpretable and realistic social knowledge prior would allow for 1) flexibility in defining the structured knowledge present, 2) fine-grained analysis of a model's decision-making process, and 3) commonsense facts to help in modeling real-world social intelligence.

As steps towards this long-term vision, I plan to first focus on the simpler speaker-listener, non-interactive setting where a human speaker conveys their beliefs and intentions via verbal and nonverbal behaviors to the listener. While such settings are well studied through modeling human affective states, personalities, and cognitive states, existing research primarily focuses on task-specific supervised approaches. As a result, they do not model interpretable knowledge priors of general social commonsense. I aim to incrementally construct such knowledge via three stages: 1) injecting prior knowledge by defining entities and relations inspired by research in modeling social intelligence from psychology and philosophy, 2) a data-driven approach that leverages large banks of unlabeled single-speaker data and multimodal feature extractors, and 3) integrating both classical and learned entities and relations. The resulting knowledge graph should represent the beliefs of the listener as they infer traits displayed by the speaker. While predictive tasks such as emotion recognition might primarily rely on additive signals, I am more interested in studying tasks involving sarcasm, humor, and deception where information from modalities is often contradictory. I plan to evaluate whether social commonsense knowledge can improve sample complexity, interpretability, and controllability beyond purely data-driven supervised methods. Relying partially on interpretable, predefined knowledge can also help in more robust and fair learning.

4.4 Real-world Applications

4.4.1 Ongoing work in learning markers of suicide from mobile data: Suicide is the second leading cause of death among adolescents, with 16% of high school students reporting seriously considering suicide each year, and 8% making one or more suicide attempts [4]. As a step towards adaptive interventions of suicidal behaviors, intensive monitoring of behavior via adolescents' use of smartphones may shed new light on the early risk of suicidal thoughts and behaviors. While smartphones provide a valuable data source, one must take care to summarize behaviors from mobile data without identifying the user through personal (e.g., personally identifiable information) or protected attributes (e.g., race, gender). As part of my broader research in real-world social AI, I am currently designing algorithms to **learn multimodal privacy-preserving markers of suicidal thoughts and behaviors from mobile data**, with the modalities spanning indicators of emotional distress (acoustic voice data, communicative language, facial expression, and music choice), social dysfunction (content and patterns of online communication, geographic movement), and sleep disturbance (actigraphy, light sensors, and diurnal patterns of phone use). I have made progress in fair and computationally-efficient federated learning algorithms for decentralized multimodal device data [12, 16]. My current goal is to achieve a balance between predictive performance of STBs and protecting the privacy of personal and protected attributes.

5 Conclusion

Long-term vision: I believe that social AI can democratize access in areas of beneficial social impact such as improving the quality of healthcare and maximizing the accessibility of education. Therefore, I envision a world with synergy between humans and AI connected via social interaction, where intelligent robots can engage with humans and offer assistance in schools, hospitals, and the workplace. My vision of accessible AI also leads me to focus on ensuring the robustness and fairness of such models so that no social group will be at a disadvantage when deployed. While this is an ambitious goal, I believe that my expertise in multimodal learning and close collaboration with wonderfully gifted researchers in academia and industry places me in a perfect position to execute this research plan.

References

[1] Anushri Arora, Akanksha Joshi, Kruttika Jain, Shashank Dokania, and Pravin Srinath. Unraveling depression using machine intelligence. In *ICCES*. IEEE, 2018.

[2] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *NIPS*, 2016.

[3] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette Chang, Sungbok Lee, and Shrikanth S. Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Journal of Language Resources and Evaluation*, 42(4):335–359, dec 2008.

[4] CDC. *Suicide Facts at a Glance 2015*, 2015 (accessed September 6, 2020).

[5] Cheng Chen, Qi Dou, Yueming Jin, Hao Chen, Jing Qin, and Pheng-Ann Heng. Robust multimodal brain tumor segmentation via feature disentanglement and gated fusion. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 447–456. Springer, 2019.

[6] Ning Gao, Xingyuan Wang, and Xiukun Wang. Multi-layer progressive face alignment by integrating global match and local refinement. *Applied Sciences*, 9(5):977, 2019.

[7] Angeliki Lazaridou, Anna Potapenko, and Olivier Tieleman. Multi-agent communication meets natural language: Synergies between functional and structural language learning. In *ACL*, 2020.

[8] **Paul Pu Liang**, Ziyin Liu, AmirAli Bagher Zadeh, and Louis-Philippe Morency. Multimodal language analysis with recurrent multistage fusion. In *EMNLP*, 2018.

[9] **Paul Pu Liang**, Ruslan Salakhutdinov, and Louis-Philippe Morency. Computational modeling of human multimodal language: The mosei dataset and interpretable dynamic fusion. *Carnegie Mellon University*, 2018.

[10] **Paul Pu Liang**, Amir Zadeh, and Louis-Philippe Morency. Multimodal local-global ranking fusion for emotion recognition. In *ICMI*, 2018.

[11] **Paul Pu Liang**, Yao Chong Lim, Yao-Hung Hubert Tsai, Ruslan Salakhutdinov, and Louis-Philippe Morency. Strong and simple baselines for multimodal utterance embeddings. In *NAACL-HLT*, 2019.

[12] **Paul Pu Liang**, Terrance Liu, Liu Ziyin, Nicholas B. Allen, Randy P. Auerbach, David Brent, Ruslan Salakhutdinov, and Louis-Philippe Morency. Think locally, act globally: Personalized and efficient federated learning with local and global representations. *NeurIPS Workshop on Federated Learning*, 2019.

[13] **Paul Pu Liang**, Zhun Liu, Yao-Hung Hubert Tsai, Qibin Zhao, Ruslan Salakhutdinov, and Louis-Philippe Morency. Learning representations from imperfect time series data via tensor rank regularization. In *ACL*, 2019.

[14] **Paul Pu Liang**, Jeffrey Chen, Ruslan Salakhutdinov, Louis-Philippe Morency, and Satwik Kottur. On emergent communication in competitive multi-agent teams. In *AAMAS*, 2020.

[15] **Paul Pu Liang**, Irene Li, Emily Zheng, Yao Chong Lim, Ruslan Salakhutdinov, and Louis-Philippe Morency. Towards debiasing sentence representations. In *ACL*, 2020.

[16] **Paul Pu Liang**, Terrance Liu, Anna Cai, Michal Muszynski, Ryo Ishii, Nicholas Allen, Randy Auerbach, David Brent, , Ruslan Salakhutdinov, and Louis-Philippe Morency. Learning language and multimodal privacy-preserving markers of mood from mobile data. *ACL*, 2021.

[17] **Paul Pu Liang**, Peter Wu, Liu Ziyin, Louis-Philippe Morency, and Ruslan Salakhutdinov. Cross-modal generalization: Learning in low resource modalities via meta-alignment. *ACM Multimedia*, 2021.

[18] Louis-Philippe Morency, Rada Mihalcea, and Payal Doshi. Towards multimodal sentiment analysis: Harvesting opinions from the web. In *Proceedings of the 13th International Conference on Multimodal Interfaces*, 2011.

[19] Eda Okur, Shachi H Kumar, Saurav Sahay, and Lama Nachman. Audio-visual understanding of passenger intents for in-cabin conversational agents. *arXiv preprint arXiv:2007.03876*, 2020.

[20] Sunghyun Park, Han Suk Shim, Moitreya Chatterjee, Kenji Sagae, and Louis-Philippe Morency. Computational analysis of persuasiveness in social multimedia: A novel dataset and multimodal prediction approach. In *ICMI*. ACM, 2014.

[21] Hai Pham*, **Paul Pu Liang***, Thomas Manzini, Louis-Philippe Morency, and Barnabás Póczos. Found in translation: Learning robust joint representations by cyclic translations between modalities. In *AAAI*, 2019.

[22] Yao-Hung Hubert Tsai*, **Paul Pu Liang***, Amir Zadeh, Louis-Philippe Morency, and Ruslan Salakhutdinov. Learning factorized multimodal representations. In *ICLR*, 2019.

[23] Xuejiao Yang and Bang Wang. Local ranking and global fusion for personalized recommendation. *Applied Soft Computing*, page 106636, 2020.

[24] AmirAli Bagher Zadeh, **Paul Pu Liang**, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *ACL*, 2018.