Social Influence Dialogue Systems for Social Good

Current dialogue systems are primarily for information seeking or social companionship. However, they fail to proactively apply strategies in complex and critical social influence tasks, such as persuading people to perform physical exercise or providing personalized counseling. **My research grants dialogue systems social influence abilities to promote good social causes like human experts. Key themes of my work include developing techniques for intelligible dialogue generation and privacy protection to make such systems deployable in real life.**

- **Social influence dialogues** employ strategies to influence users' attitudes or behavior. Such dialogues span various domains including persuasion and recommendation. I proposed *PersuasionForGood*, a new persuasive donation task. It received a ***best paper nomination*** at ACL 2019 [1], a top-tier conference, and has been widely adopted in NLP research [2,3]. I study the social influence dynamics of persuasive donation, which lays the groundwork for personalized persuasive dialogue systems. My work also examines users' perceptions of AI agent identities [4]. This influential study validates the necessity of Autobot Law across the nation, cautions against the misuse of chatbot identities, proposes regulations on social influence system design, and could lead to the enactment of related legislation. I was also a core member of the team behind a ***Science* publication** [5], where we built the first human-level negotiation AI agent, *Cicero*, in the well-known strategy board game of *Diplomacy* that involves natural language cooperation, negotiation, and persuasion between seven players [6].
- **Intelligible dialogue generation.** Existing chatbots frequently repeat or contradict themselves (e.g., *"I have never heard of the charity. It is my favorite charity!"*), which impedes the social influence process. A traditional solution is to train a supervised classifier to detect and filter out incoherent dialogue generations. However, my work allows dialogue generation models to introspect and identify their own mistakes without an extra classifier [7]. It is a radically new approach with both strong performance and small annotation efforts. I apply it to the aforementioned *Diplomacy* game to improve the AI negotiation agent, and it achieves the same state-of-the-art result of a large supervised classifier.
- **Privacy Protection.** Conversations often sparsely include personal information. Inspired by this observation, I proposed a new notion–*Selective Differential Privacy* (SDP)–to protect selected user private information in language data. My work also includes effective privacy mechanisms to achieve robust SDP-protected models [8,9]. It is one of the pioneering studies in the space of privacy-preserving NLP models and has inspired multiple new research directions in the community [10, 11].

In summary, I develop intelligible and safe dialogue systems for social influence and am recognized as a **Rising Star in Machine Learning** for related studies. So far, our persuasion projects have raised $2000+ donations to a charity called *Save the Children* to help children all over the world. Meanwhile, the proposed methods are broadly applicable to general dialogue systems and different learning and inference tasks. Besides the focus on social influence dialogue systems, my research is interdisciplinary, connecting natural language processing (NLP) with social science, human-computer interaction (HCI), and cybersecurity. I am always excited to collaborate with researchers from different fields.

1 Social influence dialogue systems

Social influence is ubiquitous in life [12], in situations ranging from healthy habit promotion to emotional support. My research tackles two main challenges around social influence dialogue systems: 1) how to model user social identity to personalize such systems, and 2) how perceptions of AI identities impact the social influence outcome. These methods and results can also apply to general dialogue systems.

Persuasive dialogue system for donation. I proposed a novel persuasion task for social good, where one participant was asked to persuade the other to donate to a children's charity [1]. Proposing this task involved developing a rich persuasive dialogue dataset with user personality and persuasive strategy annotation. Our analysis on the data shows that strategies have different effects on different users: for instance, *emotional appeal* is more effective for extroverted people. This work laid the groundwork for personalized persuasive dialogue systems and inspired new directions such as emotion-aware [2] and persona-aware [3] persuasive dialogues. My follow-up study [13] built a persuasive dialogue system with

reinforcement learning and imitation learning, and achieved a 70% increase in donation over human persuaders. I have also studied other aspects of social influence, such as dialogue systems for movie recommendations [14] and scam prevention [15].

The impact of chatbot identity on social influence. In 2019, California proposed the Autobot Law [16], which was the first to require businesses to disclose chatbot identities. At that time, little was known about how chatbot identities would impact conversational outcomes, especially in the context of social influence. To answer this question, we conducted an online factorial experiment [4] with hidden and disclosed chatbot identities on the donation persuasion task. We found that people are more likely to donate money when they *think* they are talking to other humans, which proves the necessity of the Autobot Law across the country. In cases where humans are aware that they are speaking with a chatbot, they are more likely to donate if the chatbot is more competent. This suggests that improving dialogue quality is crucial for successful social influence outcomes. This is one of the first works to caution against the misuse of chatbot identity and guide social influence dialogue system design, which could promote the enactment of legislation in related areas.

2 Intelligible dialogue generation

Successful dialogue systems must be competent. This is especially true for social influence tasks, as smooth user experiences are essential to successful social influence. However, existing dialogue agents often produce incoherent utterances, including repetitive and contradictory statements, which greatly hurts the user experience. My research towards intelligible dialogue generation addresses the following questions: 1) how to detect unintelligible messages, and 2) how to generate intelligible messages.

Unintelligible message detection. Current methods adopt supervised methods to detect unintelligible messages and filter them out afterwards, but this requires an external classifier in addition to the dialogue generation model in a dialogue system. My work enables dialogue generation models to identify their own mistakes introspectively without another classifier. Intuitively, if a generation model predicts that the user is more likely to respond *"I don't understand"* to a message it generated, then this generated message may be unintelligible. We propose a novel algorithm to search for such discriminative continuations following unintelligible messages, and use their probabilities to detect incoherence in a semi-supervised fashion without a separate classifier. My approach matches the state-of-the-art performance of a carefully fine-tuned large supervised classifier in the complex negotiation dialogues from the game of *Diplomacy* [7].

Intelligible message generation. Once we can detect nonsensical messages, the next step is to improve dialogue models themselves. Previous work employed reinforcement learning (RL) to improve dialogue models by interacting with a sophisticated user simulator [17]. My work refines dialogue models without any simulator [13]: the model first explores the space by generating multiple message candidates, then identifies the good and bad candidates, and finally learns from its own mistakes via negative rewards. After each turn, the conversation continues with the original trajectory without a user simulator. My approach improved the dialogue quality over state-of-the-art baselines by 15% in the persuasion task and received positive user feedback.

3 Privacy protection

Through my research in social influence dialogue systems, I realize that people often share personal information in conversations, such as their name and address, creating concerns about user privacy. Differential privacy (DP) [18] is a dominant privacy notion for privacy protection. The key idea of DP is to carefully inject noise into the learning algorithm so that the model does not rely too much on any training example. However, traditional DP learning algorithms protect the entire training example and thus suffer from low utility when only partial information in an example is sensitive. For instance, in NLP applications, if the user mentions *"My zip code is 12345"*, only the tokens *"12345"* with the actual zip code need to be protected.

New privacy notion for NLP – Selective DP. To improve the private model utility in NLP, my work formalizes an effective new privacy notion–*Selective Differential Privacy* (SDP)– that protects sensitive portions of language data [8, 9]. To realize SDP, I have developed privacy mechanisms tailored for

different model architectures: 1) *Selective-DPSGD* for RNN-based models, and 2) *JFT* for transformer-based models. Experiments show that our mechanisms achieve better model utilities while remaining safe under different privacy attacks compared to state-of-the-art approaches. This work has inspired many related studies in privacy-preserving NLP [10, 11]. Moreover, despite our focus on NLP, SDP could be useful for other tasks where partial data are sensitive, such as face recognition in computer vision.

Protecting missed secrets. SDP protects the sensitive information identified by any secret detector (a model that detects private information). One pressing concern in SDP is the privacy leakage of secrets missed by an imperfect secret detector. My work is the first to systematically protect these missed secrets with both empirical techniques and theoretical analyses [9]. Since the portion of missed secrets is small, intuitively, we need smaller noise to ensure their privacy. To compute the needed small noise, we estimate secret detectors' missing rate M, leverage *privacy amplification by subsampling* [19] to calculate the privacy parameters associated with M, and apply a private optimizer with the calculated small noise to fine-tune the model to achieve SDP. Experiments show that our approach improves the model utility while protecting the missed secrets from empirical attacks in tasks beyond dialogue systems.

4 Related studies in task-oriented and open-domain dialogues

My dialogue research has also developed novel methods in related topics such as task-oriented and open-domain dialogues, and dialogue evaluation. Towards intelligible dialogue generation, I have designed task-oriented systems to adapt to user sentiment [20], studied user simulators for RL-based systems [17], and developed algorithms to extract dialogue structure under low-resource settings [21]. Towards more sociable systems, I have proposed methods to integrate user feedback [22], which will be used to improve BlenderBot 3 [23], a social chatbot with millions of users. Moreover, I have developed a toolkit [24] for easy dialogue evaluation and deployment, which has been adopted in various projects.

5 Future directions

In sum, I build social influence dialogue systems that can interact with humans naturally and responsibly. Moving forward, I am passionate about the following socially impactful and practical problems.

- **Multi-party dialogue systems.** If we deploy persuasive agents on a large scale, enabling it to interact with multiple users simultaneously will make the conversations more engaging, realistic and robust. Besides, multi-party conversations are common in real life, such as business meetings and in-class discussions. Therefore, I plan to study multi-party dialogue systems and start with these problems: 1) how to understand who is speaking to whom, 2) how to decide when and whom to talk to, and 3) how to track these dialogue states.
- **Ethical dialogue systems.** I am passionate about building well-intentioned systems for marginalized groups in this technology-driven world, e.g., to accompany the elderly, educate the young, and counsel the ones in need. However, marginalized groups are often underrepresented in studies, e.g., only 2.1% of participants in our study are senior citizens. So the first step towards ethical dialogue systems is to invite various underrepresented groups for studies and understand their user dynamics to break the technology barriers for them.
- **Learning through interactions.** Dialogue agents interact with users and thus should evolve with human feedback. This involves three future research problems: 1) how to update the models offline with collected feedback, 2) how to adjust the dialogue trajectories online given real-time reactions, and 3) how to interleave these two steps towards systems that can continuously evolve.
- **Democratizing AI.** Alongside the learning through interaction efforts, I am also interested in building a community, where everyone can build their own ML models for various tasks, and provide feedback to help improve others' models. In this way, we can involve the general public in AI development and educate them about AI technologies to democratize AI and benefit society.

My vision is to build a natural interface between human intelligence and machine intelligence via dialogues, which can be used to gather human feedback and improve different applications, from automatic code generation to interactive robot learning. With such a natural interface, all members of society can interact with AI models seamlessly and benefit from AI advances.

References

[1] Xuewei Wang*, **Weiyan Shi***, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. Persuasion for good: Towards a personalized persuasive dialogue system for social good. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019.

[2] Sara Asai, Koichiro Yoshino, Seitaro Shinagawa, Sakriani Sakti, and Satoshi Nakamura. Emotional speech corpus for persuasive dialogue system. In *Proceedings of The 12th Language Resources and Evaluation Conference*, 2020.

[3] Abhisek Tiwari, Tulika Saha, Sriparna Saha, Shubhashis Sengupta, Anutosh Maitra, Roshni Ramnani, and Pushpak Bhattacharyya. A persona aware persuasive dialogue policy for dynamic and co-operative goal setting. *Expert Systems with Applications*, 2022.

[4] **Weiyan Shi**, Xuewei Wang, Yoo Jung Oh, Jingwen Zhang, Saurav Sahay, and Zhou Yu. Effects of persuasive dialogues: testing bot identities and inquiry strategies. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI)*, 2020.

[5] FAIR, Anton Bakhtin*, Noam Brown*, Emily Dinan*, Gabriele Farina, Colin Flaherty*, Daniel Fried, Andrew Goff, Jonathan Gray*, Hengyuan Hu*, Athul Paul Jacob*, Mojtaba Komeili, Karthik Konath, Minae Kwon, Adam Lerer*, Mike Lewis*, Alexander H. Miller*, Sasha Mitts, Adithya Renduchintala*, Stephen Roller, Dirk Rowe, **Weiyan Shi***, Joe Spisak, Alexander Wei, David Wu*, Hugh Zhang*, and Markus Zijlstra. Human-Level Play in the Game of Diplomacy by Combining Language Models with Strategic Reasoning. *Science*, 2022. *A core contributor of the team. Authors listed alphabetically.

[6] Allan Dafoe, Yoram Bachrach, Gillian Hadfield, Eric Horvitz, Kate Larson, and Thore Graepel. Cooperative AI: Machines must learn to find common ground. *Nature*, 593(7857):33–36, 2021.

[7] **Weiyan Shi**, Emily Dinan, Adi Renduchintala, Daniel Fried, Athul Paul Jacob, Zhou Yu, and Mike Lewis. AutoReply: Detecting Nonsense in Dialogue Introspectively with Discriminative Replies. *Under Submission*, 2022.

[8] **Weiyan Shi**, Aiqi Cui, Evan Li, Ruoxi Jia, and Zhou Yu. Selective Differential Privacy for Language Modeling. *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2022.

[9] **Weiyan Shi**, Ryan Shea, Si Chen, Chiyuan Zhang, Ruoxi Jia, and Zhou Yu. Just Fine-tune Twice: Selective Differential Privacy for Large Language Models. *Empirical Methods in Natural Language Processing (EMNLP)*, 2022.

[10] Xuandong Zhao, Lei Li, and Yu-Xiang Wang. Provably Confidential Language Modelling. *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2022.

[11] Antonio Ginart, Laurens van der Maaten, James Zou, and Chuan Guo. Submix: Practical private prediction for large-scale language models. *arXiv preprint arXiv:2201.00971*, 2022.

[12] Keise Izuma. The neural bases of social influence on valuation and behavior. In *Decision Neuroscience*. Elsevier, 2017.

[13] **Weiyan Shi**, Yu Li, Saurav Sahay, and Zhou Yu. Refine and Imitate: Reducing Repetition and Inconsistency in Persuasion Dialogues via Reinforcement Learning and Human Demonstration. *Findings of the Association for Computational Linguistics: EMNLP 2021*, 2021.

[14] Shirley Anugrah Hayati, Dongyeop Kang, Qingxiaoyang Zhu, **Weiyan Shi**, and Zhou Yu. INSPIRED: Toward sociable recommendation dialog systems. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.

[15] Yu Li, Kun Qian, **Weiyan Shi**, and Zhou Yu. End-to-end trainable non-collaborative dialog system. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2020.

[16] California Governor. California new Autobot Law, Cal. Bus. & Prof. Code § 17940, et seq. (SB 1001), 2018.

[17] **Weiyan Shi***, Kun Qian*, Xuewei Wang, and Zhou Yu. How to build user simulators to train rl-based dialog systems. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019.

[18] Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 2014.

[19] Borja Balle, Gilles Barthe, and Marco Gaboardi. Privacy amplification by subsampling: Tight analyses via couplings and divergences. *Advances in Neural Information Processing Systems*, 2018.

[20] **Weiyan Shi** and Zhou Yu. Sentiment adaptive end-to-end dialog systems. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2018.

[21] **Weiyan Shi**, Tiancheng Zhao, and Zhou Yu. Unsupervised Dialog Structure Learning. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2019.

[22] **Weiyan Shi**, Emily Dinan, Kurt Shuster, Jason Weston*, and Jing Xu*. When Life Gives You Lemons, Make Cherryade: Converting Feedback from Bad Responses into Good Labels. *arXiv preprint arXiv:2210.15893*, 2022.

[23] Kurt Shuster, Jing Xu, Mojtaba Komeili, Da Ju, Eric Michael Smith, Stephen Roller, Megan Ung, et al. BlenderBot 3: a deployed conversational agent that continually learns to responsibly engage. *arXiv preprint arXiv:2208.03188*, 2022.

[24] Yu Li, Josh Arnold, Feifan Yan, **Weiyan Shi**, and Zhou Yu. Legoeval: An open-source toolkit for dialogue system evaluation via crowdsourcing. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations (ACL Demo)*, 2021.