Social Influence Dialogue Systems

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. **My research grants dialogue systems social influence abilities to promote good social causes. Key themes of my work include 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 [20], a top-tier conference, and has been widely adopted in NLP research. 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 [17]. This seminal work cautions against the misuse of chatbot identities and proposes guidance on regulating social influence system design.

- **Intelligible dialogue generation**. Existing chatbots frequently repeat or contradict themselves (e.g., *"I have never heard of the charity. They are my favorite charity!"*), which impedes the social influence process. A standard solution is to train a supervised classifier to detect and filter unintelligible dialogue generations, but my work allows dialogue models to introspect and identify their own mistakes without an external classifier [13]. I apply my approach to the well-known negotiation board game of *Diplomacy* [7] and achieve the same state-of-the-art performance of a large supervised classifier.

- **Privacy Protection**. Conversations often sparsely include personal information. To protect user privacy in language data, I proposed a new notion–*Selective Differential Privacy* (SDP)–to protect the sensitive portions of conversations. My work also includes effective privacy mechanisms to achieve robust SDP-protected models [11, 12]. 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 [4, 21].

In summary, I develop intelligible and safe dialogue systems for social influence. As described above, my research is highly interdisciplinary, connecting natural language processing (NLP) with various fields such as social science, human-computer interaction (HCI), and cybersecurity. I am always excited to collaborate with researchers from different fields.

1 Social influence dialogue systems

Most dialogue studies focus on task-oriented or open-domain dialogue agents for information-seeking tasks or social companionship. But dialogues that influence users' attitudes or behavior with strategies are equally important. My research tackles two main challenges around social influence dialogue systems: 1) how to build such systems, and 2) how perceptions of AI identities impact the social influence outcome.
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 [20]. The task came with a rich persuasive dialogue dataset with user personality and persuasion strategy annotation. Our analysis 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 [1] and persona-aware [10] persuasive dialogues. My follow-up study [14] imitates human persuasion strategies and achieves a 70% increase over human persuaders. I have also studied other aspects of social influence, such as dialogue systems for movie recommendations [6].
The impact of chatbot identity on social influence. In 2019, California proposed the Autobot Law [5], which was the first to require businesses to disclose chatbot identities. At the 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 [17] 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 validates 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 pioneering works to caution against the misuse of chatbot identity and guide social influence dialogue system design.

2 Intelligible dialogue generation

Successful social influence systems must be competent. 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 answers the following questions: 1) how to detect nonsensical messages, and 2) how to generate intelligible messages.

Nonsensical message detection. Current methods adopt supervised methods to detect nonsensical messages, 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 believes that the user is more likely to respond *"I don't understand"* to a message it generated, then this generated message may be nonsense. We propose a novel algorithm to search for such discriminative continuations following nonsensical messages, and use their probabilities to detect nonsense in a semi-supervised fashion without a separate classifier. My approach matches the state-of-the-art performance of a carefully fine-tuned supervised classifier in the complex negotiation dialogues from the game of *Diplomacy*.

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 [15]. My work refines dialogue models without any simulator [14]: 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 dialogue systems, I realize that people often share personal information in conversation, such as their name and address, creating concerns about user privacy. Differential privacy (DP) [3] is a dominant privacy notion for privacy protection. 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 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 [11, 12]. 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 [4, 21]. 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 [11]. Since the portion of missed secrets is small, intuitively, we need smaller noise to ensure their privacy. To theoretically compute the needed small noise, we estimate secret detectors' missing rate M, leverage *privacy amplification by subsampling* [2] 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.

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

My exploration of dialogue research has also created 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 [18], studied user simulators for RL-based systems [15], and developed algorithms to extract dialogue structure under low-resource settings [19]. Towards more sociable systems, I have proposed methods to integrate user feedback [16], which will be used to improve BlenderBot 3 [9], a social chatbot with millions of users. Moreover, I have developed a toolkit [8] for easy dialogue evaluation and deployment, which has been adopted in various projects.

5 Future directions

In summary, I build social influence dialogue systems that can interact with humans naturally and responsibly. I believe that dialogue is an interface between human intelligence and machine intelligence and can be used to gather human feedback and improve various 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 the AI advancement. Moving forward, I am passionate about the following socially impactful and practical problems.

- **Multi-party dialogue systems.** If we deploy our persuasive agent on a large scale, enabling it to interact with multiple users simultaneously will make the conversations more engaging. 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.** My biggest passion is to build 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, only 2.1% of participants in our study are senior citizens. So my 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, similar to Wikipedia, where everyone can build their own ML models for various tasks, and provide feedback to help edit 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.

References

[1] 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*, pages 491–497, 2020. URL: https://aclanthology.org/2020.lrec-1.62/.

[2] Borja Balle, Gilles Barthe, and Marco Gaboardi. Privacy amplification by subsampling: Tight analyses via couplings and divergences. *Advances in Neural Information Processing Systems*, 2018. URL: https://dl.acm.org/doi/10.5555/3327345.3327525.

[3] Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 2014. URL: https://dl.acm.org/doi/10.1561/0400000042.

[4] 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. URL: https://openreview.net/pdf?id=cKTBRHIVjy9.

[5] California Governor. California new autobot law, cal. bus. & prof. code § 17940, et seq. (sb 1001), 2018. https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201720180SB1001.

[6] 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. URL: https://aclanthology.org/2020.emnlp-main.654.

[7] Dave de Jonge, Tim Baarslag, Reyhan Aydoğan, Catholijn Jonker, Katsuhide Fujita, and Takayuki Ito. The challenge of negotiation in the game of diplomacy. In *International Conference on Agreement Technologies*. Springer, 2018. URL: https://link.springer.com/chapter/10.1007/978-3-030-17294-7_8.

[8] 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. URL: https://aclanthology.org/2021.acl-demo.38.

[9] Kurt Shuster, Jing Xu, Mojtaba Komeili, Da Ju, Eric Michael Smith, Stephen Roller, Megan Ung, Moya Chen, Kushal Arora, Joshua Lane, et al. Blenderbot 3: a deployed conversational agent that continually learns to responsibly engage. *arXiv preprint arXiv:2208.03188*, 2022. URL, https://arxiv.org/abs/2208.03188.

[10] 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. URL: https://dl.acm.org/doi/abs/10.1016/j.eswa.2021.116303.

[11] **Weiyan Shi**, Si Chen, Chiyuan Zhang, Ruoxi Jia, and Zhou Yu. Just fine-tune twice: Selective differential privacy for large language models. *arXiv preprint arXiv:2204.07667*, 2022. URL: https://arxiv.org/abs/2204.07667.

[12] **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: Human Language Technologies (NAACL)*, 2022. URL: https://aclanthology.org/2022.naacl-main.205.

[13] **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. URL: https://openreview.net/forum?id=adtvwY0UcNH.

[14] **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. URL: https://aclanthology.org/2021.findings-emnlp.295/.

[15] **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. URL: https://aclanthology.org/D19-1206.

[16] **Weiyan Shi**, Kurt Shuster, Emily Dinan, Jing Xu, and Jason Weston. When life gives you lemons, make cherryade: Converting feedback from bad responses into good labels. *under submission*, 2022.

[17] **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. URL: https://dl.acm.org/doi/10.1145/3313831.3376843.

[18] **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. URL: https://aclanthology.org/P18-1140.

[19] **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: Human Language Technologies (NAACL)*, 2019. URL: https://aclanthology.org/N19-1178.

[20] 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. URL: https://aclanthology.org/P19-1566/.

[21] 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: Human Language Technologies (NAACL)*, 2022. URL: https://aclanthology.org/2022.naacl-main.69/.