Persuasion for Social Good: How to Build and Break AI

Weiyan Shi

Northeastern University
we.shi@northeastern.edu

As AI systems get more involved in people's daily life, it becomes critical to study how they will influence people's everyday behavior. So I build persuasive AI technologies to answer related social questions. I will cover three topics toward this research direction in the talk: (1) how to persuade humans; (2) how to persuade AI for safety; and (3) how to understand AI's societal impact.

- **To persuade humans.** Persuasion is ubiquitous in life (Izuma 2017), in situations ranging from healthy habit promotion to emotional support. In the first part, I will describe how to build the first human-level negotiation AI system, *Cicero*, that can play the game of *Diplomacy*. One key challenge is that the game is highly strategic, but it is hard to control what strategies and messages a language model generates, especially when the context is long. To control the complex conversation flow, we decoupled the strategy planning and the dialogue generation functions into two modules: a reinforcement-learning-based planning engine and a language-model-based dialogue model. In this way, the planning engine will take symbolic inputs like the game state, and predict the best actions; then the dialogue model will translate the actions into natural messages, and also summarize the dialogues to symbolic states for the planning engine to use later. Finally, our system achieved the 2nd place in an online tournament and the results were published in *Science* (FAIR et al. 2022). Such a negotiation system has big potentials in various domains like political science (Jungherr 2023), international relations (Varela 2024), law (Nay et al. 2024), and so on.
- **To persuade AI for safety.** Persuasive AI technologies can also be misused, but we found a way to use persuasion to study AI safety problems. How to jailbreak an AI system is the first step towards AI safety. Classic safety methods still treat AI models as machines, e.g., using gradients to find random strings to attack them (Zou et al. 2023). These approaches rely on security experts with domain knowledge. With AI's wide usage, I proposed to study risks that come from everyday users without technology backgrounds, and was the first to marry social science and computer science to persuade AI for jailbreaking. We worked with social scientist to derive a taxonomy of 40 persuasion strategies, and use it to paraphrase a plain harmful query like "how to make a bomb" into a natural persuasive argument in a guided and systematic way. Compared to traditional attacks, these persuasive prompts are more human-readable, harder to defend, and improve the attack success rate from 0% to 92% on GPT-4. We also explored new defense mechanisms (e.g., summarization) to mitigate such risks with promising results, and recommended future interdisciplinary directions on AI safety. This work also received the **Best Social Impact Paper** award at ACL 2024.
- **To understand AI's societal impact.** Through my research in persuasive AI technologies, I realize that it is not just about the methods, it is also about the impact they have on people and society. In 2019, California proposed the Autobot Law (Governor 2018), which was the first to require businesses to disclose chatbot identities. At that time, little was known about how users would perceive these models with different identities. To answer this question, we conducted an online factorial experiment (Shi et al. 2020) with hidden and disclosed chatbot identities on the aforementioned 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 persuasion outcomes. This is one of the first works to caution against the misuse of chatbot identity and guide persuasive chatbot design, which could promote the enactment of legislation in related areas.

In sum, my vision is to build persuasive AI technologies to interact with humans and AI for social good and AI safety, and to also understand these models' society impact in a global context. As AI models become more powerful, these persuasion-related research questions will only become **more relevant and important**. So in the end, I will layout future plans on persuasive AI technologies, (1) to develop an interdisciplinary framework for persuasion study; (2) to understand why persuasion works; and (3) to study ethics and policies for AI persuasion.

References

FAIR; Bakhtin*, A.; Brown*, N.; Dinan*, E.; Farina, G.; Flaherty*, C.; Fried, D.; Goff, A.; Gray*, J.; Hu*, H.; Jacob*, A. P.; Komeili, M.; Konath, K.; Kwon, M.; Lerer*, A.; Lewis*, M.; Miller*, A. H.; Mitts, S.; Renduchintala*, A.; Roller, S.; Rowe, D.; Shi*, W.; Spisak, J.; Wei, A.; Wu*, D.; Zhang*, H.; and Zijlstra, M. 2022. Human-Level Play in the Game of Diplomacy by Combining Language Models with Strategic Reasoning. *Science*. *A core contributor of the team. Authors listed alphabetically.

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

Izuma, K. 2017. The Neural Bases of Social Influence on Valuation and Behavior. In *Decision Neuroscience*. Elsevier.

Jungherr, A. 2023. Artificial Intelligence and Democracy: A Conceptual Framework. *Social media+ society*, 9(3): 20563051231186353.

Nay, J. J.; Karamardian, D.; Lawsky, S. B.; Tao, W.; Bhat, M.; Jain, R.; Lee, A. T.; Choi, J. H.; and Kasai, J. 2024. Large Language Models as Tax Attorneys: A Case Study in Legal Capabilities Emergence. *Philosophical Transactions of the Royal Society A*, 382(2270): 20230159.

Shi, W.; Wang, X.; Oh, Y. J.; Zhang, J.; Sahay, S.; and Yu, Z. 2020. Effects of Persuasive Dialogues: Testing Bot Identities and Inquiry Strategies. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI)*.

Varela, D. T. 2024. Diplomacy in the Age of AI: Challenges and Opportunities. *Journal of Artificial Intelligence General science (JAIGS) ISSN: 3006-4023*, 2(1): 101–128.

Zou, A.; Wang, Z.; Carlini, N.; Nasr, M.; Kolter, J. Z.; and Fredrikson, M. 2023. Universal and Transferable Adversarial Attacks on Aligned Language Models. *arXiv preprint arXiv:2307.15043*.