Statement of Purpose

Zen (Zhengxuan Wu), CS Ph.D. Applicant, Fall 2020

Introduction

From Google Home to Tesla Autopilot, artificial intelligence (AI) is progressing rapidly. To build the next generation of artificial general intelligence (AGI), I believe that it is essential for AIs to have human-like learning abilities and cognitive awareness of surroundings. In this way, AIs can collaborate with humans; that is, they can understand our intentions and reward systems as individuals (as in human-computer interaction and inverse reinforcement learning[1–3]), our learning and hierarchical reasoning abilities (as in probabilistic modeling[4–6]), our ability to attribute mental states to others (as in theory of mind[7–9]), and our societal networks with interpersonal collaborations and information propagation (as in computational social science[10–12]). These advanced capabilities will foster a more powerful AI system, allowing it to learn from limited data, make reliable predictions, and accomplish smooth and naturalistic interactions with humans.

Inspired by these research areas, it is my goal to pursue a Ph.D. in computer science at Stanford University, with a focus on bridging social and cognitive science with AI systems. Specifically, I am interested in (1) building cognitively inspired AI agents to collaborate intelligently with human and other agents and (2) using AI tools to study social cognition. Bolstered by my experience developing strong engineering skills at VMware and research institutions including Case Western Reserve, University of Pennsylvania, and Stanford, as well as my solid academic background in cognitive science, psychology, and social sciences, I am fortunate to possess the practical experience, technical knowledge, and interdisciplinary perspective needed to approach problems in fields bridging cognitive science and AI.

Experience

During my graduate study at Stanford with a focus on computational social science, I was fortunate to work with researchers in psychology, cognitive science, social science, and computer science. Within my broad research spectrum, all of my experiences centered on reasoning about humans by using computational models.

Reasoning of Intuitive Psychology Under the guidance of Dr. Desmond Ong and Dr. Jamil Zaki at Stanford, my current research focuses on enabling AI systems to have human-like cognition of emotion. Our topics include (1) predicting emotional states of others with multi-modal inputs, (2) reverse engineering how humans intuitively reason about other people from trained parameters and comparing that with brain activities, and (3) codifying such reasoning via probabilistic modeling: a human-like approach that involves both symbolic encoding for knowledge representations and hierarchical reasoning using Bayesian inferences. This extended journey has culminated in the publication of our Stanford Emotional Narratives Dataset on IEEE Transaction on Aective Computing, as well as the building of a Transformer-based memory fusion network model and variational neural networks to accurately predict emotional states of humans, at the Aective Computing Intelligent Interaction Conference with me as first author.[13,14] Currently, we are continuing our research in building asynchronized multi-stream LSTMs and recurrent multi-modal VAEs to predict and infer emotional state with missing modalities. Additionally, we are extending our work to reasoning of intuitive physics. By understanding human cognition, I am confident that researchers can build AI agents that are capable of building robots that understand the internal states of others, and are fair and safe while interacting with humans and other agents.[15,16]

Reasoning of Human Behaviors Through computational models with proper design and training through persuasive systems, we can better understand the human cognition process, which can further be used to influence human perception and behavior collaboratively.[17,18] With this in mind, I was fortunate to work on HabitLab, a project led by Dr. Geza Kovacs and Dr. Michael Bernstein at Stanford. HabitLab is a chrome browser extension that contains a variety of productivity interventions aiming to reduce the time spent on user-specified websites or applications. Leveraging in-the-wild experiences with online interventions, we used this platform to study and influence user behaviors in a naturalistic way. I supported these efforts by helping with outlining the interplay between efficacy and intervention attrition rates. To optimize efficacy, I also helped in building adaptive interventions that are optimized for individuals using a multi-armed bandit algorithm. We also looked at the conservation of procrastination across multiple devices. Specifically, we investigated whether productivity interventions can actually help users to save time, or just redistribute it across devices. Additionally, we investigated changes in user motivation over time, as observed through the lens of intervention difficulty levels, and submitted our paper to CHI2020. Through this study, I have contributed to two papers that were published in CSCW 2018 and CHI 2019 as full papers.[19,20] Through this valuable experience, I learned how AIs can not only better understand our behaviors and meet our needs, but also cooperate with humans to augment our intelligence.

Reasoning of Learning Enabling an AI system to learn like a child remains a major unsolved problem.[21] Revealing the underlying learning mechanisms in humans can help us to better construct human-like AI agents. With this research question in mind, I had the opportunity to work with Dr. Michal Kosinski and Dr. Poruz Khambatta at Stanford, leading a new study about whether humans can learn how to judge other peoples political views from faces, how they learn, and how they learn differently from computer models (Demo). During the experiment, participants invited to the lab were shown face images of politicians, and then asked about the political views of the people in the image, while receiving incentives for getting correct answers and penalties for wrong ones. We used an augmented Q-learning model to impute the learning rate for each participant. Additionally, we are collecting personal traits to compare learning rates across various groups with different traits. Furthermore, we plan to collect eye-tracking data to compare the attended regions on faces of humans and deep learning models. This experience has helped us to characterize the human learning process, which in turn has helped us to build models that can learn like humans.

Reasoning of Social Psychology To discover how AIs and computational models in general can help humans to understand cognition beyond the individual level, I was fortunate to work with Dr. Michal Kosinski in the field of computational social science, endeavoring to understand social cognition using digital footprints. Digital footprints can predict social traits, which include sexual orientation.[22] Our study has focused on the differences in social traits between heterosexuals and homosexuals, using data mining and deep learning methods on Facebook datasets from the myPersonality website. We concluded that masculinity-femininity scores can predict sexual orientation in males but not in females. We illustrated how our work aligns with previous psychology experiments that elucidate the power of digital footprints in studying social cognition,[23] and are close to submission of paper to a psychology journal paper.[24] From these studies, I discovered the power of digital footprints in understanding human social traits, but also became aware that this power comes at the cost of privacy.[22] It is with this philosophy in mind that I would like to build AI systems that are intelligent and safe.

Besides these experiences, I also participated in projects related to probing semantics of emotions with word embedding,[25] modeling interpersonal emotion differentiation,[26] modeling facial movements that imply emotions,[27] reasoning of morality, and analyzing social networks of Chinese politicians,[28] which all led to publications and manuscripts in preparation. Each of these challenging yet rewarding projects has broadened my research spectrum, but also affirmed that I want to work on building AI systems with human-like cognition and ability to learn.

Interests

If admitted to Stanford, I would be especially keen to work with professors affiliated with the *Stanford Institute for Human-Centered Artificial Intelligence*. Specifically, I would be honored to work with Prof. Jiajun Wu, Prof. Noah Goodman and Prof. Daniel Yamins (cognitive AI), Prof. Michael Bernstein and Prof. Dorsa Sadigh (human and computer collaborations), and Prof. Jure Leskovec (computational social science). The specific topics in which I am interested range from applications to the theoretical level and from deep learning to generative modeling, but all of them involve human reasoning.

One research topic in which I am interested is reverse engineering current deep learning models and building cognitively inspired AI models. I am eager to improve deep learning and other learning paradigms to move closer to human-like learning by incorporating psychological ingredients. One example would be automated statistical reasoning techniques such as program induction or probabilistic programming approaches to solve problems in the context of intuitive physics and intuitive psychology (as in works by Prof. Jiajun Wu and Prof. Noah Goodman[29–31]). I would be interested in working to interpret human cognition of morality or emotion, based on deep learning models and probabilistic programming. Another topic that excites me is building interpretable Bayesian models to learn causality relations in intuitive physics, and comparing computer models with brain activities (as in works by Prof. Daniel Yamins[32,33]). I believe that future AI systems should be capable of few-shot learning by extracting causal relationships from limited samples, as humans can do.

Another related topic that excites me is human-computer and multi-agent interactions. I would be interested in building an end-to-end human-computer interaction system with automated cognitive models that understand human needs and cooperates with humans to achieve goals (as in the work of Prof. Michael Bernstein[34,35]). In the process, I would be interested in drawing insights from Bayesian models of human cognition, so that machines could infer humans internal emotions and moral states.[36,37] This maps well to multi-agent probabilistic planning, as in order to have computer systems interact well with users, computer programs should better infer the cognitive state of users over time. Additionally, I would be keen to study cooperative and multi-agent learning for a better understanding of human behavior, to improve human-robot interaction, and to facilitate cooperation in multi-agent systems with different goals and states (as in the work of Prof. Dorsa Sadigh[38,39]).

Social cognition is another topic in which I am interested. I would be keen to work on computational social science, including developing network-based models and deep learning models to study psychological traits of individuals and crowd behavior on a large scale, whether these be physical, biological, social, or ethical (as in the work of Prof. Jure Leskovec[40]). I am interested to engage with deep representation learning in networks and multimodal learning based on knowledge graphs. I am also keen to apply computational models to large-scale data, the web, and online media to study social cognition and diffusion of information.

Above all, it is my goal to undertake research that promises to make a safe, robust, and reliable difference for our foreseeable future. I deeply believe that future AI systems should be human-centric. My robust interdisciplinary perspective gained from basic physics, computer science, and social science prepares me to contribute to the community in the CS program. In turn, I believe the rich research community at Stanford will provide me with invaluable training in both social science and computer science, along with a cohort of exceptional peers. Thank you for your consideration.

References

[1] Dorsa Sadigh, S Shankar Sastry, Sanjit A Seshia, and Anca Dragan. Information gathering actions over human internal state. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 66–73. IEEE, 2016.

[2] Andrew Y Ng, Stuart J Russell, et al. Algorithms for inverse reinforcement learning. In *Icml*, volume 1, page 2, 2000.

[3] Dylan Hadfield-Menell, Stuart J Russell, Pieter Abbeel, and Anca Dragan. Cooperative inverse reinforcement learning. In *Advances in neural information processing systems*, pages 3909–3917, 2016.

[4] Andreas Stuhlmüller and Noah D Goodman. Reasoning about reasoning by nested conditioning: Modeling theory of mind with probabilistic programs. *Cognitive Systems Research*, 28:80–99, 2014.

[5] Nick Chater, Joshua B Tenenbaum, and Alan Yuille. Probabilistic models of cognition: Conceptual foundations, 2006.

[6] Nick Chater and Christopher D Manning. Probabilistic models of language processing and acquisition. *Trends in cognitive sciences*, 10(7):335–344, 2006.

[7] Henry M Wellman. *The child's theory of mind.* The MIT Press, 1992.

[8] Josef Perner. *Understanding the representational mind.* The MIT Press, 1991.

[9] David Premack and Guy Woodruff. Does the chimpanzee have a theory of mind? *Behavioral and brain sciences*, 1(4):515–526, 1978.

[10] David Lazer, Alex Pentland, Lada Adamic, Sinan Aral, Albert-László Barabási, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, et al. Computational social science. *Science*, 323(5915):721–723, 2009.

[11] John Scott. Social network analysis. *Sociology*, 22(1):109–127, 1988.

[12] Stanley Wasserman, Katherine Faust, et al. *Social network analysis: Methods and applications*, volume 8. Cambridge university press, 1994.

[13] Desmond C Ong, Zhengxuan Wu, Tan Zhi-Xuan, Marianne Reddan, Isabella Kahhale, Alison Mattek, and Jamil Zaki. Modeling emotion in complex stories: the Stanford Emotional Narratives Dataset. *IEEE Transactions on Affective Computing*, to appear.

[14] Zhengxuan Wu, Xiyu Zhang, Tan Zhi-Xuan, Jamil Zaki, and Desmond C. Ong. Attending to emotional narratives. *IEEE Affective Computing and Intelligent Interaction (ACII)*, 2019.

[15] Andrea Bajcsy, Dylan P Losey, Marcia K O'Malley, and Anca D Dragan. Learning from physical human corrections, one feature at a time. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, pages 141–149. ACM, 2018.

[16] Andrea Bajcsy, Dylan P Losey, Marcia K OMalley, and Anca D Dragan. Learning robot objectives from physical human interaction. *Proceedings of Machine Learning Research*, 78:217–226, 2017.

[17] Brian J Fogg. Persuasive technology: using computers to change what we think and do. *Ubiquity*, 2002(December):5, 2002.

[18] Harri Oinas-Kukkonen and Marja Harjumaa. Persuasive systems design: Key issues, process model, and system features. *Communications of the Association for Information Systems*, 24(1):28, 2009.

[19] Geza Kovacs, Zhengxuan Wu, and Michael S Bernstein. Rotating online behavior change interventions increases effectiveness but also increases attrition. *Proc. ACM Hum.-Comput. Interact.*, 2(CSCW), November 2018.

[20] Geza Kovacs, Drew Mylander Gregory, Zilin Ma, Zhengxuan Wu, Golrokh Emami, Jacob Ray, and Michael S Bernstein. Conservation of procrastination: Do productivity interventions save time or just redistribute it? In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, page 330. ACM, 2019.

[21] Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. Building machines that learn and think like people. *Behavioral and brain sciences*, 40, 2017.

[22] Michal Kosinski, David Stillwell, and Thore Graepel. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, 110(15):5802–5805, 2013.

[23] J Richard Udry and Kim Chantala. Masculinity–femininity predicts sexual orientation in men but not in women. *Journal of Biosocial Science*, 38(6):797–809, 2006.

[24] Zhengxuan Wu and Michal Kosinski. Homosexual women are not masculine, 2019.

[25] Zhengxuan Wu and Yueyi Jiang. Disentangling latent emotions of word embeddings on complex emotional narratives. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 587–595. Springer, 2019.

[26] Erik Nook, Christina Chwyl, Isabella Kahhale, Zhengxuan Wu, and Jamil Zaki. Interpersonal emotion differentiation, 2019.

[27] Alison Mattek, Michael Smith, Zhengxuan Wu, Isabella Kahhale, Marianne Reddan, Desmond Ong, and Jamil Zaki. Modeling facial movements that track emotion inference, 2019.

[28] Yueyi Jiang, Zhengxuan Wu, Arseny Ryazanov, and Piotr Winkielman. Social influence shifts gamble preferences for monetary and moral decisions, 2019.

[29] Jiajun Wu, Ilker Yildirim, Joseph J Lim, Bill Freeman, and Josh Tenenbaum. Galileo: Perceiving physical object properties by integrating a physics engine with deep learning. In *Advances in neural information processing systems*, pages 127–135, 2015.

[30] Noah D Goodman, Joshua B Tenenbaum, and Tobias Gerstenberg. Concepts in a probabilistic language of thought. Technical report, Center for Brains, Minds and Machines (CBMM), 2014.

[31] Desmond Ong, Harold Soh, Jamil Zaki, and Noah Goodman. Applying probabilistic programming to affective computing. *IEEE Transactions on Affective Computing*, 2019.

[32] Charles F Cadieu, Ha Hong, Dan Yamins, Nicolas Pinto, Najib J Majaj, and James J DiCarlo. The neural representation benchmark and its evaluation on brain and machine. *arXiv preprint arXiv:1301.3530*, 2013.

[33] Damian Mrowca, Chengxu Zhuang, Elias Wang, Nick Haber, Li F Fei-Fei, Josh Tenenbaum, and Daniel L Yamins. Flexible neural representation for physics prediction. In *Advances in Neural Information Processing Systems*, pages 8799–8810, 2018.

[34] Sharon Zhou, Tong Mu, Karan Goel, Michael Bernstein, and Emma Brunskill. Shared autonomy for an interactive ai system. In *The 31st Annual ACM Symposium on User Interface Software and Technology Adjunct Proceedings*, pages 20–22. ACM, 2018.

[35] Melissa A Valentine, Daniela Retelny, Alexandra To, Negar Rahmati, Tulsee Doshi, and Michael S Bernstein. Flash organizations: Crowdsourcing complex work by structuring crowds as organizations. In *Proceedings of the 2017 CHI conference on human factors in computing systems*, pages 3523–3537. ACM, 2017.

[36] Thomas M Moerland, Joost Broekens, and Catholijn M Jonker. Emotion in reinforcement learning agents and robots: a survey. *Machine Learning*, 107(2):443–480, 2018.

[37] Chao Yu, Minjie Zhang, Fenghui Ren, and Guozhen Tan. Emotional multiagent reinforcement learning in spatial social dilemmas. *IEEE transactions on neural networks and learning systems*, 26(12):3083–3096, 2015.

[38] Dorsa Sadigh, Shankar Sastry, Sanjit A Seshia, and Anca D Dragan. Planning for autonomous cars that leverage effects on human actions. In *Robotics: Science and Systems*, volume 2. Ann Arbor, MI, USA, 2016.

[39] Dorsa Sadigh, Katherine Driggs-Campbell, Alberto Puggelli, Wenchao Li, Victor Shia, Ruzena Bajcsy, Alberto Sangiovanni-Vincentelli, S Shankar Sastry, and Sanjit Seshia. Data-driven probabilistic modeling and verification of human driver behavior. In *2014 AAAI Spring Symposium Series*, 2014.

[40] Jure Leskovec and Andrej Krevl. Snap datasets: Stanford large network dataset collection, 2014.