11-866 Artificial Social Intelligence **Spring 2023**

Week 10: Embodiment and Non-Verbal Generation

Instructors: Louis-Philippe Morency, Paul Liang *Synopsis Leads: Soham D. Tiwari, Rosa Vitiello*

Edited by Paul Liang *Scribes: Siyu Chen, Jiaxin Shi*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

Follow the rest of the class here: https://cmu-multicomp-lab.github.io/asi-course/spring2023/

Summary: In week 10's discussion session, the class continued to explore literature in artificial social agents. In contrast to our discussion a week prior on conversational AI, this week focused on the importance of embodiment and non-verbal generation in AI, as well as the challenges and tradeoffs associated in this field of research.

The following was a list of provided research probes:

1. Last week, we studied recent technologies to model written-text conversations. This week, we study spoken and embodied interactions, which also include nonverbal communication. What are the new technical challenges that emerge when building spoken and embodied conversational AI?
2. Large language models have shown impressive performances for text-based conversational AI. Why did we not see the same improvements for embodied conversational AI? Is it just a question of more data? Better annotations? Or should nonverbal and multimodal interactions be modeled differently?
3. Take a moment to review the papers from week 5 which listed methods to evaluate social agents, including proposed taxonomies for social skills and social intelligence. Do current models and systems properly address the abilities and skills of social intelligence? How should we build new technologies for better embodied social agents?
4. As we start thinking beyond dyadic interactions and dialogues, what are the elements of social intelligence that are particularly important in small group settings? What are the new technical challenges to build social agents in these group settings? Can you think of concepts and technologies originally built for dyadic interactions which will not generalize to the group settings?
5. One extensively-discussed problem is about the type of embodiment. For example, should the social agent have physical embodiment (e.g., robots) or virtual (e.g., virtual humans). Think about the impact of embodiment on the problem of building social agents, more specifically its social skills and social intelligence. What are the differences? Which ones will generalize between embodiments?
6. As a small extra exercise, imagine a world without transformers or large-scale language models (i.e., Vaswami et al. paper was never published). How would you start addressing this problem of building embodied social agents if these tools (transformer, LLMs) were not available? In other words: take a fresh look at the problem of artificial social intelligence, without grounding it too much on current models and systems. How would you solve the problem? Where to start?

As background, students read the following papers:

1. (Required) Social Robotics [Breazeal et al., 2016]
2. (Required) The Fabric of Socially Interactive Agents: Multimodal Interaction Architectures [Kopp and Hassan, 2022]
3. (Suggested) A Comprehensive Review of Data-Driven Co-Speech Gesture Generation [Nyatsanga et al., 2023]
4. (Suggested) Learning to Listen: Modeling Non-Deterministic Dyadic Facial Motion [Ng et al., 2022]
5. (Suggested) Interaction in Social Space [Vilhjálmsson, 2022]
6. (Suggested) Core Challenges in Embodied Vision-Language Planning [Francis et al., 2021]

7. (Other) A Survey of Embodied AI: From Simulators to Research Tasks [Duan et al., 2021]
8. (Other) Real Robots, Real Learning Problems [Brooks and Matarić, 1993]

We summarize several main takeaway messages from group discussions below:

1 Challenges in Developing Spoken and Embodied AI

Several new challenges must be overcome when developing spoken and embodied agents with non-verbal skills. The challenges are broadly related to perception, the generation of non-verbal signals, and manipulating the robot in a physical environment.

Moravec's paradox is an observation by AI and robotics researchers that AI easily does tasks like complex computation and reasoning, which are difficult for humans. On the other hand, perception and sensorimotor capabilities are very challenging for AI and require a lot of computing [Jeevanandam, 2022]. Consequently, it is challenging to develop support for visual and speech perception by embodied agents, which is out of scope and hence not a concern for text-only conversational AI. The slow verbal communication and instantaneous non-verbal signals occur over different time durations. Hence the agent will have to parallelly process language, emotions, and gestures while grounding them in the cultural context of the social environment.

Similar to the problems faced in perception, the model will also need to be capable of generating various non-verbal signals in different timescales. Moreover, special care would need to be taken that the nonverbal signals generated by the agent are related to the utterances being spoken (which operate on a longer timescale compared to instantaneous gestures). In addition, the embodied agent must also be able to make gestures, which entail nuanced movement of various body parts. This leads to a need for good control theory models for nuanced and fine-grained control of the different actuators in such an embodied agent. Finally, while the embodied agent moves around and manipulates objects in the physical world, it must also be cognizant of proxemics and understand a comfortable level of closeness to the user during interaction.

The technical challenges accompanying the development of the above features range from the accumulation of appropriate data to learning models of fine-grained motion. First, to develop such embodied and spoken AI agents, data must be collected to support the learning of multimodal social perception and social generation. This data would then be further used for training the agents.

The models used for training such agents would need special considerations for a large generation space and robustness in situations of uncertainty. There are numerous possible combinations of non-verbal facial expressions, gestures, textual content, and speech tonality, and it would be very challenging to model such a large joint output space. Consequently, some assumptions would have to be made to optimize generation in the large joint output space, or the joint output space would have to be broken down into discrete output spaces. Furthermore, in uncertain unprecedented social situations, the model must be robust enough to generate some output that suits the situation or have special error-handling messages without crashing the system to maintain user trust.

2 LLMs and Embodied Conversational AI

To address themes in Probe 2, we discussed the impact of large language models in textual conversational AI and why it has not yet made the same impact in embodied applications. We highlight various reasons why large language models techniques have not been as widely popular in embodied conversational AI.

A contributing reason why LLMs and the growth of conversational AI has not been seen in embodied AI is due to its complexity. Embodied AI includes a wide range of modalities which introduces the integration and alignment problem commonly faced in multimodal tasks. Moreover, these range of modalities also brings low-level signal challenges, such as perception in video data, as well as higher computation costs. The requirements of rich input and output for embodied AI makes collecting data and designing effective annotations more costly.

The complexity of challenges in embodied tasks has likely contributed to slower research growth in the field; however, it continues to be a promising research path as new data, methods, and better computation are developed. In fact, recent advances have been made in introducing LLMs to embodied AI, most notably Google's embodied multimodal language model Palm-E and its success in a variety of embodied reasoning tasks [Driess et al., 2023].

3 Social Skills in Embodied ASI

Following themes from Probe 3, we theorized how to implement a couple social skills in embodied ASI:

- **Mirroring:** In social interactions, humans often mirror each other when conversing [Pineda, 2009]. While these skills often happen naturally in humans, some studies focus on impacts of teaching autistic populations how to mirror or appropriately converse [Vivanti and Rogers, 2014]. Using these techniques based on theory from cognitive science, we may also be able to teach ASI similar social skills.
- **Memory:** Memory is a large part of how humans interact socially and with others. For example, remembering previous interactions or preferences while conversing with others. While models do have the ability to memorize, often their memory is not as long-term as humans. For long-term memory, current commercial ASI products, such as Moxie, cache certain data about users to create a personalized profile for those interacting with the agent [Hurst et al., 2020]. In shorter-term, memory may be kept using a short window of interaction history (i.e., a set window of prior time steps), which may be sufficient for certain tasks, such as social listening [Ng et al., 2022].

4 Social AI for Group Interactions

Several challenges arise for artificial social agents when the social interaction settings are scaled up from dyadic to a group. The challenges are related to multiple speakers speaking simultaneously. There is a possibility that there could be multiple parallel threads of communication in a group. So the first challenge is speaker diarization, i.e., the ability to map the ASR transcriptions to the different speakers. The next challenge will be to model and distinguish between communication threads if multiple subgroups talk simultaneously. Finally, most groups today are quite diverse, hence the need for the agent to model the differences in the group participants (cultural, accent, etc.) and personalize the responses using appropriate grounding when speaking to a particular group member.

5 Physical vs. Simulated Embodied AI

Social artificially intelligent agents can be embodied physically or virtually (via a graphical representation only.) One of the reasons to choose physical embodiment over virtual is that previous works show that humans trust physically embodied agents more than virtually embodied agents [Thellman et al., 2016]. Moreover, the physical embodiment enables greater functionality in task completion capabilities and the potential for more natural communications with users. However, a disadvantage of a physical embodiment is that it requires the model to process another input modality and the need to develop robust control theory models to produce realistic gestures while taking all measures to ensure the physical movements of the robot do not harm anybody. This is where virtual embodied agents are more useful, as they do not need to be concerned about the physical aspects, which then translates to faster development types for the prototypes and fewer points of failure. Consequently, virtual embodied agents could easily adapt to various environments. Hence the choice between virtual and physical embodied agents would largely depend on the end-use case of the agent.

6 Modeling ASI without Transformers

In our group discussions, we imagined a world where Vaswani et al. [2017] was never published. We discussed why transformers became popular in 2017 and whether they are necessary to develop competent ASI.

Alternatives to Transformers Below we summarize discussions on alternative architectures to transformers:

- **Modular approaches:** Instead of end-to-end transformers, we may break down tasks in a modular-fashion, priortizing a subtask type architecture. This modular design may be more interpretable and effective in task-oriented applications.
- **Knowledge or rule-based approaches:** A top-down rule-based method may also be effective in specific tasks. This is typical in robotics applications, such as those used by Boston Dynamics Robotics. These methods have clear interpretability as it can be seen when certain rules are applied to a given situation. In order for rule-based approaches to be effective, a detailed taxonomy and lexicons of all possible non verbal signals, different categories, conscious and subconscious emotional states, body movements, different vocal non-verbal signals is likely needed. In order to develop this, insights from domain knowledge of social psychology may be useful.
- **Other end-to-end architectures:** While transformers are popular, other model architectures may be applicable, such as CNNs or RNNs that are shown to deal with sequential data relatively effectively and at lower computational cost than transformers.

Are Transformers Necessary? We also discussed whether we need transformers and whether all situations require transformers as a solution. Transformers prioritize linearlization and how to pay attention to output. It is important to note that these models are highly effective, particularly in tasks with large amounts of data. In certain tasks, transformers may not be desired. For example, tasks with less data may not be sufficient to train an effective transformer model depending on the extent of few-shot or zero-shot learning. Moreover, it may be unnecessary to use a transformer in specific tasks where rule-based or modular methods are sufficient and may be more efficient

7 Ethical Concerns in Embodied AI

In both group discussions, we considered ethical concerns of designing embodied AI. As embodied social agents become vital tools in everyday life, it is crucial to acknowledge and weigh the potential positive and negative impacts of this technology. We summarize potential risks of embodied AI, as well as thoughts on how to design robots to keep in mind ethical considerations.

General ethical and safety standards in social AI As robotics and embodied AI have become more popular, there has been a growing interest in establishing ethical standards towards the development of social agents [Winfield, 2019, Van Maris et al., 2020]. That is, when creating both embodied and virtual AI, it is important to consider their environmental and user impact from both a safety and ethical perspective.

For example, in contrast to virtually embodied AI, robots have the ability to physically manipulate the space and persons around them. While this allows physically embodied AI to complete and aid in a wider range of tasks, it also poses dangers to the user if they malfunction, such as when a robot broke a player's finger while playing chess [Henley, 2022].

While robots raise concerns about responsibilities of physical safety and manipulation of their environment, social agents also pose other ethical social harms, particularly in critical applications, such as social companions for elderly, mental healthcare, and education. There are also ethical concerns in how social agents and models may spread misinformation or bias which has been highlighted in the recent growing use of large language models [Weidinger et al., 2021]. When designing social agents, it is important to account for these ethical and safety standards.

Tradeoffs in personalized embodied AI In current work, there is a wave of interest in designing personalized ASI, such as embodied agents that react to users needs and preferences [Hurst et al., 2020, Park et al., 2019]. Personalized embodied AI may have many benefits, such as better user outcomes and longer term relationships and interaction. However, personalization also poses risks, such as concerns of privacy. Moreover, similar to concerns of recommendation systems, a highly personalized AI may cause over-engagement or reinforcement of biased views. With highly effective ASI, one may also use the technology to persuade users which could be used maliciously.

Designing agents with ethical considerations Below are summarized discussion points on how to design robots to take into account ethical concerns in embodied AI:

- Carefully designing objective functions that ensure conservative and safe decisions from social agents.
- To address issues in personalization of agents, include noise in user inputs so that the model does not over fit to the user.

References

Cynthia Breazeal, Kerstin Dautenhahn, and Takayuki Kanda. Social robotics. *Springer handbook of robotics*, pages 1935–1972, 2016.

Rodney A. Brooks and Maja J. Matarić. Real robots, real learning problems. 1993.

Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. Palm-e: An embodied multimodal language model. In *arXiv preprint arXiv:2303.03378*, 2023.

Jiafei Duan, Samson Yu, Tangyao Li, Huaiyu Zhu, and Cheston Tan. A survey of embodied ai: From simulators to research tasks. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 6: 230–244, 2021.

Jonathan M Francis, Nariaki Kitamura, Felix Labelle, XiaoPeng Lu, Ingrid Navarro, and Jean Oh. Core challenges in embodied vision-language planning. *J. Artif. Intell. Res.*, 74:459–515, 2021.

Jon Henley. Chess robot grabs and breaks finger of seven-year-old opponent, July 2022. URL https://www.theguardian.com/sport/2022/jul/24/chess-robot-grabs-and-breaks-finger-of-seven-year-old-opponent-moscow.

Nikki Hurst, Caitlyn E. Clabaugh, Rachel Baynes, Jeffrey F. Cohn, Donna D. Mitroff, and Stefan Scherer. Social and emotional skills training with embodied moxie. *ArXiv*, abs/2004.12962, 2020.

Dr Nivash Jeevanandam. What is true about Moravec's paradox?, 2022. URL https://indiaai.gov.in/article/what-is-true-about-moravec-s-paradox.

Stefan Kopp and Teena Hassan. *The Fabric of Socially Interactive Agents: Multimodal Interaction Architectures*, page 77–112. Association for Computing Machinery, New York, NY, USA, 1 edition, 2022. ISBN 9781450398961. URL https://doi.org/10.1145/3563659.3563664.

Evonne Ng, Hanbyul Joo, Liwen Hu, Hao Li, Trevor Darrell, Angjoo Kanazawa, and Shiry Ginosar. Learning to listen: Modeling non-deterministic dyadic facial motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20395–20405, June 2022.

Simbarashe Nyatsanga, Taras Kucherenko, Chaitanya Ahuja, Gustav Eje Henter, and Michael Neff. A comprehensive review of data-driven co-speech gesture generation. *arXiv preprint arXiv:2301.05339*, 2023.

Hae Won Park, Ishaan Grover, Samuel Spaulding, Louis Gomez, and Cynthia Breazeal. A model-free affective reinforcement learning approach to personalization of an autonomous social robot companion for early literacy education. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19. AAAI Press, 2019. ISBN 978-1-57735-809-1. doi: 10.1609/aaai.v33i01.3301687. URL https://doi.org/10.1609/aaai.v33i01.3301687.

Jaime A Pineda. *Mirror neuron systems: The role of mirroring processes in social cognition*. Springer, 2009.

Sam Thellman, Annika Silvervarg, Agneta Gulz, and Tom Ziemke. Physical vs. virtual agent embodiment and effects on social interaction. volume 10011, pages 412–415, 09 2016. ISBN 978-3-319-47664-3. doi: 10.1007/978-3-319-47665-0_44.

Anouk Van Maris, Nancy Zook, Praminda Caleb-Solly, Matthew Studley, Alan Winfield, and Sanja Dogramadzi. Designing ethical social robots—a longitudinal field study with older adults. *Frontiers in Robotics and AI*, 7:1, 2020.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Hannes Högni Vilhjálmsson. *Interaction in Social Space*, page 3–44. Association for Computing Machinery, New York, NY, USA, 1 edition, 2022. ISBN 9781450398961. URL https://doi.org/10.1145/3563659.3563662.

Giacomo Vivanti and Sally J. Rogers. Autism and the mirror neuron system: insights from learning and teaching. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369, 2014.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*, 2021.

Alan Winfield. Ethical standards in robotics and ai. *Nature Electronics*, 2(2):46–48, 2019.