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

Week 5: Artificial Social Agents - Measurements

Instructors: Louis-Philippe Morency and Paul Liang *Synopsis Leads: Leena Mathur and Ryan Liu*

Edited by Paul Liang *Scribes: Soham Tiwari and Hwijeen Ahn*

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: Artificial social intelligence (ASI) encompasses the research effort towards advancing AI systems that can perceive and interpret human social information (e.g., communication patterns, pragmatics, intent) and engage in seamless human-AI interactions. In week 5's discussion session, the class focused on discussing best practices and goals for designing effective questionnaires, user studies, and benchmarks for measuring social intelligence in virtual and embodied AI systems. The following was a list of provided probes:

1. With a view towards artificial social agents (robots, virtual humans,...), how should we measure the quality of social intelligence, social competence and social skills? Should we apply the measures from human social intelligence directly to artificial social intelligence? Where are the main differences?
2. Researchers in robotics, virtual humans and artificial intelligence started building questionnaires and benchmarks to measure social intelligence. How could we potentially improve these proposed questionnaires and benchmarks, possibly using the knowledge gained from studying human social intelligence (including human social competence and human social skills)?
3. The Tian and Oviatt paper [Tian and Oviatt, 2021] studies socio-emotional competence with a slightly different perspective, looking instead to the errors happening during human-robot interactions? Should we focus more on the errors or on the competence? How are these two concepts related? Does competence mean lack of errors? Should we try to bring competence and errors in the same framework?
4. Many of the questionnaires for social robots and virtual humans are based on direct interaction with the AI system. After the interaction, the human subject is asked to fill the questionnaire to evaluate and measure social intelligence. This is in contrast with a more "offline" approach, such as Social IQA, where scenarios are written and shared with the AI system directly, to evaluate its social intelligence.Which approach should we prioritize? What are the pros and cons of both paradigms (evaluation after direct interaction vs simulated written scenarios)? What are the other evaluation paradigms you think would be helpful to measure social intelligence of artificial social agents?
5. Do you expect social intelligence to be measured differently between an physically-embodied AI (e.g., robot) or a virtually-embodied AI (e.g., virtual human)? What about non-embodied agents or agents with different embodiment (e.g., HAL 9000)? How should we measure social intelligence?

As background, students read the following papers:

1. A Taxonomy of Social Errors in Human-Robot Interaction [Tian and Oviatt, 2021]
2. Measuring the Perceived Social Intelligence of Robots [Barchard et al., 2020]
3. The Artificial-social-agent Questionnaire: Establishing the Long and Short Questionnaire Versions [Fitrianie et al., 2022b]
4. SOCIAL IQA: Commonsense Reasoning about Social Interactions [Sap et al., 2019]
5. Artificial Social Agent Questionnaire Instrument [Fitrianie et al., 2022a]
6. The 19 Unifying Questionnaire Constructs of Artificial Social Agents: An IVA Community Analysis [Fitrianie et al., 2020]
7. Practical Intelligence [Hedlund, 2020]

We summarize main takeaway messages from group discussions below, organized by discussion probe:

1 Measuring Social Intelligence, Competency, Skills

Differences among measuring social intelligence, competencies, and skills: When measuring the quality of social intelligence, social competence, and social skills, our group characterized their differences along the dimension of *concreteness*. On one hand, social competence and skills are context-specific; therefore, they can be concretely measured in strictly-controlled environments that enable the construction of reproducible contexts (e.g. experiments in a simulation environment such as AI2-Thor [Kolve et al., 2017]). Social intelligence is more abstract and latent than competencies and skills, as established by [Weis and Conzelmann, 2015]. Therefore, it may not be feasible to directly measure ”social intelligence” – it would make sense to estimate social intelligence via an aggregation of questionnaires and texts spanning different contexts.

1.1 Measurement methods

We discussed model-agnostic approaches, where we evaluate all agents using the same base metric regardless of their implementation or embodiment. One benefit of this approach is that the end user may not know the inner workings of the model; therefore, this type of test is aligned with user interpretability. However, we should note that there is an inherent trade-off between interpretability and feasibility towards achieving a model's goals, so evaluating models using implementation-specific methods should also be considered afterwards.

Confounding Factors: It is especially important to consider confounding factors in experiments to assess social intelligence, competencies, and skills. Such confounders include the age, identity, and cultural norms of annotators, as well as the participant data being annotated. Of these, the *identity* confounder is particularly of interest as it can be influenced by the agent's role in the interaction. Even within the same culture, social relationships and interactions can be *polyperceivable*, ranging from objective (e.g. recognizing a ”father-son” relationship) to subjective (e.g. “friends”).

Subjectivity in Measurement: On the note of subjectivity, an agent's capabilities could alternatively be measured by *perceived competence*, focusing on the user experience of the agent's behaviors. This digs at a more foundational question: Can we make an agent seem socially competent in different contexts, without it necessarily being socially intelligent? Furthermore, humans can sometimes also manage their incorrect urges, and others are none the wiser as long as downstream actions are not affected. Drawing from this as inspiration, it may be unnecessary for us to evaluate the intermediate variables of a model, further motivating a “black box”, model-agnostic approach. In most modern methods, the self-regularization or post-hoc judgement system is already incorporated as part of the model structure as well (e.g. selecting one label after the final layer of a model is inherently a best selection over classes).

1.2 Transferring Evaluation Methods of Social Intelligence between Contexts

One of the problems we focused on is how to transfer the measurement of social intelligence across contexts. The context could differ among tasks, modalities, or model architecture and output.

For the aforementioned model-agnostic approaches to evaluation, the previous methods can still be used if the tasks and output domains of the new context are roughly similar. However, we also consider when the output space changes completely. Under these scenarios, we highlight a type of approach that breaks the task into sub-components. One example of this is FILM: Following Instructions in Language with Modular methods [Min et al., 2021], where authors decompose a task into understand and execute stages.

Another issue in evaluation when we have a gold-standard method (or human solutions) is when there is a mismatch is procedural steps. An example would be when the human solution has two discrete steps represented in language, but the robot solution domain has three. One heuristic we could use to bridge this gap is to jointly optimize for reducing the agent's number of instruction steps. However, this might be undesirable when instructions are not sufficiently clear. Another option is to mapping one task to multiple tasks to try to force a match between the data.

Lastly, it is worth mentioning that if the model that already performs well can evaluate a wide input space (e.g. an LLM), then we could simply evaluate the new model's outputs with the old model. However, this is a niche case that is not generalizable at the current stage of ML models. We note that since we are dealing with social agents, procedural explanations, reasoning, and transparency are as important as being right. Thus, a comprehensive evaluation suite would take these into account.

2 Improving Questionnaires and Benchmarks

We categorize the agent decision process into two phases, perception and decision-making, each with potential errors. Perception here entails perceiving human state during collaborations with humans, and is necessary but not sufficient for social intelligence. Furthermore, the term "perceive" is different between robot agents and human experience. Robots perceive stimuli uniformly at the lowest level, and their model architecture has the entire data space to work with to determine the output. Humans subconsciously assign importance to what they are perceiving, and a lot of modalities and portions of modalities are filtered out without ever being "perceived" as part of the human experience. Relating to what an agent experiences, another issue brought up was that context-based metrics and evaluation under context may require an evaluation approach friendlier to subjectivity, like in Tian and Oviatt [2021].

In decision-making, deterministic behavior is important for transparency and human expectation, but robots also need to be flexible. However, the degree of transparency vs. flexibility varies depending on the context: clinical assistants need to be more predictable, while artistic assistants should be more varied between outputs.

There is also a difference between the frequency distribution of errors and the relative consequences each error has, and errors are often hard to quantify properly. For instance, respecting personal space is normally important in the American social setting, but in an emergency situation is no longer appropriate. Rules are important, but knowing when to bend them, and evaluating these cases properly, is also important.

There is a similar parallel that appears in AI inference: In some cases such as emergencies, inferring what is going on is absolutely vital. However, in other cases, inference could involve stereotyping and biased judgments. And a simple solution where we mark "protected" attributes is never perfect: skincare and makeup rely on skin color, which is highly correlated with race.

Another problem is the evaluation of long tails. Exceptionally out of distribution actions may need to be penalized more harshly (e.g. slurs), but they may also be harmless (e.g. dancing on the street). Here, how to identify when to assign heavier losses is a challenging problem.

Perhaps, one could try to hardcode naive solutions such as removing slurs from the vocabulary, but then subtle micro-aggressions and other context-dependent behaviors such as harassment would not be addressed. One interesting solution suggested by the group is an iterative expanding of scenarios to fill in training examples that may be missing from rarer examples. However, this does not address the challenge of identifying an agent's problematic behavior in the first place.

Lastly, when designing questionnaires for humans to evaluate social intelligences and competencies of agents, we distinguish two important confounders: participant incentives and bias-aware question design. For the former, participants may not have aligned incentives to provide useful feedback in user study questionnaires. For example, participants in MTurk or in-person studies for money may not provide useful or extensive feedback, whereas 5-year-olds in may provide frank and extensive feedback (one group member shared an extensive experience observing this phenomena in human-robot interaction user studies). For the latter, it is important for users to not be able to guess what the purpose of the study is based on its contents, as this may invalidate the studied effect.

3 Relationships Between Errors and Competencies

Our group agreed that social errors and competencies are important to jointly study, to advance ASI.

Studying social errors in ASI systems can drive research towards improved social competencies: Identifying social errors from the user perspective can reveal new, understudied social competencies that users expect from artificial agents. One example provided was that of ”charismatic agents” – if a human stops interacting with a social robot because they find some of its gestures to be ”uncharismatic”, then that would be a social error that leads the designer to consider another social competency of ”charisma” that was not previously considered to be important.

Social competencies can be used to define goals: Studying social competencies or lack of competencies in systems can be used to define new metrics and goals for these systems. Studying social competencies can also motivate a new direction of modeling approaches (e.g., SimSensei [DeVault et al., 2014] initially expressed lots of emotions, but users did not react well to its backchanneling approaches mirroring them, leading researchers to make the models express more neutral behavior). There may also be an understudied research direction to explore on models that estimate how mistakes/social errors will impact user perceptions of social competencies.

Potential for risk-averse systems to emerge: Minimizing social errors may create risk-averse robots that are not socially-accepted or socially-approachable. Humans make mistakes frequently in interactions. Below a certain threshold, robots will be seen as more approachable if they make some minor mistakes in social interactions (e.g., pretending to get the answer wrong when tutoring a child, so the child identifies with the robot during the learning process). Vulnerable robots that make errors and admit their mistakes can lead to positive conversational dynamics during human-robot interactions [Traeger et al., 2020].

4 Online vs Offline Evaluation of Social Intelligence

Interactions with humans in the loop in real time are all considered online, while offline examples are scripted and designed to test particular areas of problems. Some more differences between online and offline are that offline evaluations can design for more control, expected responses are better defined, and exception scenarios can be tested with little risk. Online settings are much noisier, and might involve more convoluted combinations of factors.

Offline evaluations can serve as rapid benchmarks: Offline evaluations such as Social-IQa [Sap et al., 2019] can serve as rapid, static benchmark tests for basic social intelligence skills of a given agent. These offline evaluations are easy experiments to run in a controlled setting, making them reproducible and fast/smooth evaluations.

Offline evaluations can miss richness of social interactions: Benchmarks such as Social-IQa can be quite limited in how richly they capture social information in the world, as well as the dynamics of social interaction (since these benchmarks are offline/static, so will not have an interaction component)

Both online and offline evaluations can miss messiness of real-world data: Real-world data from virtual agents and physical robots interacting with humans is much messier than any online or offline NLP studies, to date. Furthermore, it is very difficult to give a score to an online experience, though Delphi is a model that succeeds at this. One could maybe consider building an LLM over experiences.

A hierarchical evaluation system: One framework that was discussed split evaluations into a three-tiered system. The lowest level is offline unit tests, testing each attribute of social skills; The intermediate level is offline integration tests, testing when multiple structured events happen at once; And the highest level is online testing, where various events can happen in the real world context. Here, the online test can be viewed as a series of integration tests, each at a different timestep.

Within this framework, an agent only needs to pass tests equivalent to its usage environment. For example, if an agent has only simple, goal-oriented tasks, offline integration tests would suffice. However, if the task is more open-ended, online testing would be needed.

5 Physical vs Virtual Embodied Social Intelligence

Our group expected social intelligence to be measured differently in physically-embodied agents (e.g., robots) and virtually-embodied agents (e.g., virtual humans), as well as between those two categories and unembodied agents (e.g., chatbot). Compared to virtually-embodied agents and unembodied agents, Physically-embodied agents have a larger space of errors, which spans physical malfunctions (e.g., robot failing to move its arms) and social errors discussed extensively in [Tian and Oviatt, 2021]. Physical errors must be carefully avoided in order to maintain user trust in robots [Rossi et al., 2017]. Our group also discussed how physically-embodied agents have more room for variations in subjective interpretation of their social and functional competence by users [Dennler et al.], since they exist in a 3D environment with a more expressive interaction space.

References

Kimberly A Barchard, Leiszle Lapping-Carr, R Shane Westfall, Andrea Fink-Armold, Santosh Balajee Banisetty, and David Feil-Seifer. Measuring the perceived social intelligence of robots. *ACM Transactions on Human-Robot Interaction (THRI)*, 9(4):1–29, 2020.

Nathaniel Dennler, Changxiao Ruan, Jessica Hadiwijoyo, Brenna Chen, Stefanos Nikolaidis, and Maja Matarić. Design metaphors for understanding user expectations of socially interactive robot embodiments. *ACM Transactions on Human-Robot Interaction*.

David DeVault, Ron Artstein, Grace Benn, Teresa Dey, Ed Fast, Alesia Gainer, Kallirroi Georgila, Jon Gratch, Arno Hartholt, Margaux Lhommet, et al. Simsensei kiosk: A virtual human interviewer for healthcare decision support. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, pages 1061–1068, 2014.

Siska Fitrianie, Merijn Bruijnes, Deborah Richards, Andrea Bönsch, and Willem-Paul Brinkman. The 19 unifying questionnaire constructs of artificial social agents: An iva community analysis. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*, pages 1–8, 2020.

Siska Fitrianie, Merijn Bruijnes, Fengxiang Li, Amal Abdulrahman, and Willem-Paul Brinkman. Artificial Social Agent Questionnaire Instrument. 11 2022a. doi: 10.4121/19650846.v3. URL https://data.4tu.nl/articles/dataset/Artificial_Social_Agent_Questionnaire_Instrument/19650846.

Siska Fitrianie, Merijn Bruijnes, Fengxiang Li, Amal Abdulrahman, and Willem-Paul Brinkman. The artificial-social-agent questionnaire: establishing the long and short questionnaire versions. In *Proceedings of the 22nd ACM International Conference on Intelligent Virtual Agents*, pages 1–8, 2022b.

Jennifer Hedlund. Practical intelligence. 2020.

Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Daniel Gordon, Yuke Zhu, Abhinav Gupta, and Ali Farhadi. Ai2-thor: An interactive 3d environment for visual ai. *arXiv preprint arXiv:1712.05474*, 2017.

So Yeon Min, Devendra Singh Chaplot, Pradeep Ravikumar, Yonatan Bisk, and Ruslan Salakhutdinov. FILM: following instructions in language with modular methods. *CoRR*, abs/2110.07342, 2021. URL https://arxiv.org/abs/2110.07342.

Alessandra Rossi, Kerstin Dautenhahn, Kheng Lee Koay, and Michael L Walters. How the timing and magnitude of robot errors influence peoples' trust of robots in an emergency scenario. In *Social Robotics: 9th International Conference, ICSR 2017, Tsukuba, Japan, November 22-24, 2017, Proceedings 9*, pages 42–52. Springer, 2017.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. Socialiqa: Commonsense reasoning about social interactions. *arXiv preprint arXiv:1904.09728*, 2019.

Leimin Tian and Sharon Oviatt. A taxonomy of social errors in human-robot interaction. *ACM Transactions on Human-Robot Interaction (THRI)*, 10(2):1–32, 2021.

Margaret L. Traeger, Sarah Strohkorb Sebo, Malte Jung, Brian Scassellati, and Nicholas A. Christakis. Vulnerable robots positively shape human conversational dynamics in a human–robot team. *Proceedings of the National Academy of Sciences*, 117(12):6370–6375, 2020. doi: 10.1073/pnas.1910402117. URL https://www.pnas.org/doi/abs/10.1073/pnas.1910402117.

Susanne Weis and Kristin Conzelmann. Social intelligence and competencies. *International Encyclopedia of the Scocial & Behavioral Sciences*, 22:371–379, 2015.