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

Week 6: Social Cognition and Social Interaction

Instructors: Louis-Philippe Morency and Paul Liang *Synopsis Leads: Rosa Vitiello, Alex Wilf*

Edited by Paul Liang *Scribes: Siyu Chen, Cathy Jiao*

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: Social cognition is a key process which humans use to understand and interact in social situations. In this week's discussion, the class aimed to define and relate social cognition to previously discussed concepts in the course, including social intelligence, social competence, and social skills. We also reflect on how these concepts can shape building of artificial social agents.

The following was a list of provided research probes:

1. How would you relate the theories from social cognition research with the concepts previously discussed in this course, including social intelligence, social skills and social competence? Can you draft an updated taxonomy (or theoretical framework) which integrates all these concepts? What are the differences between these different views?
2. "Theory of Mind" is a term that evolved over the decades, and has become popular in recent years for researchers in robotics and AI, when discussing how technologies should interact with humans. From your readings, what is missing from Theory of Mind to have a complete artificial social intelligence?
3. When reflecting on how to build artificial social agents, it is helpful to think about what are the core abilities and skills needed for AI to successfully participate in social interactions. Based on your readings and insights, what are the core elements of social interactions for artificial social agents? Lightart et al. 2021 suggests an initial list. What core elements are missing? Which elements are definitely key to social interactions?
4. Some researchers argue that social cognition research should include social interactions (aka interactionism theory). From your perspective, how much of social intelligence can be assessed offline, using tests such as the ones often used for Theory of Mind assessment? How much of social intelligence is also interactional (online) and requires different assessment methods?
5. Social perception is also an important aspect of social intelligence. The Brunswik's Lens model is a landmark model for studying human interpersonal communication. Since signals are likely to be interpreted differently, from different people, how can we model all these different perspectives of the same social interaction?

As background, students read the following papers:

1. (Required) Prospects for direct social perception: a multi-theoretical integration to further the science of social cognition [Wiltshire et al., 2015]. This paper gives a great summary of the prevalent theories and frameworks used in social cognition research. Students should focus on the first 6 pages of the paper and also pages 14-15. In other words, students should get an understanding (at least at the high-level) of the 7 theories mentioned in Figure 1, especially the first 3 theories and Brunswik's lens model.
2. (Required) Core Elements of Social Interaction for Constructive Human-Robot Interaction [Ligthart et al., 2021]. This paper brings an interesting question: what are the core components of social interactions? This paper is building on the Interactionism Theory (mentioned in the previous paper), more specifically, the paper of De Jaegher et al. 2010 (included as a Suggested Reading). While reading this paper, you should reflect on what are the core building blocks of a social interaction and, by extension, the core skills and abilities needed by an Artificial Social Agent.

3. (Suggested) Can social interaction constitute social cognition? [De Jaegher et al., 2010]. Students should focus on the first 3 pages, up to "Social interaction as contextual factor" section. While reading, be sure to understand the concepts of co-regulation, engagement and autonomy. The glossary includes an interesting definition for Social Interactions.
4. (Suggested) Social Cognition in Schizophrenia: An NIMH Workshop on Definitions, Assessment, and Research Opportunities [Green et al., 2008]. Since Social Cognition has a strong history in studying children development and some specific disorders, such Autism Spectrum Disorder and Schizophrenia, we thought it would be important to also get some definitions from the healthcare. Students should only focus on the first 2.5 pages of the paper, to get an overview of the main definitions used in Schizophrenia research, for the term social cognition.
5. (Suggested) Deconstructing and reconstructing theory of mind [Schaafsma et al., 2015]. This is an interesting opinion paper discussing the factors involved in recent Theory of Mind research, with a focus on neuroscience research. The paper categorizes typical tasks used to assess ToM. An interesting part of the paper is their attempt to factorize the problem into subcomponents (figure 3)
6. (Other) Supporting Artificial Social Intelligence With Theory of Mind [Williams et al., 2022]. This paper argues (strongly) that Theory of Mind is central to artificial social intelligence. Based on your other readings (from third week, and previous week), you should take this claim with a grain of salt. What are other components key to social intelligence, social skills and social competence? Theory of Mind is only one active theory in the field of social cognition.
7. (Other) Theory of Mind in normal development and autism [Baron-Cohen, 2001]. While this paper is a little older, it gives a concrete definition for Theory of Mind and also present many tests previously used for studying this phenomenon.
8. (Other) Enabling robotic social intelligence by engineering human social-cognitive mechanisms [Wiltshire et al., 2017]. Students should focus on the first 4 pages of the paper which summarizes major theories and frameworks in human social cognition research. This summary is similar in nature to the required reading (Wiltshire et al., 2017), but introduces the theories with a slightly different framing, which may be helpful to some readers.

We summarize several main takeaway messages from group discussions below:

1 Integrating Social Cognition & Interactions into Social Intelligence

1.1 Definitions

In the beginning of our discussion, we outlined the following definitions:

- **Social cognition:** general term that encompasses the cognition involved in determining one's own thoughts and processes (e.g., what goes on in an agent's mind, ToM)
- **Social perception:** a separate process from cognition involving what an agent perceives and how they process this information. Social perception is necessary but not sufficient for social cognition.

1.2 Creating a General Framework

In an attempt to integrate concepts from previous weeks to our understanding of social cognition, we designed a diagram to represent how social cognition connects to prior discussions, as seen in Figure 1. As depicted in the diagram, we view social cognition as a driver for social intelligence, social competence and social skills. Additionally, we find that social cognition directly affects and is affected by social perception and social interaction, as well as indirectly influenced by social expression.

We also discussed which of these components are latent versus observable. External components, such as social perception, social expression, and social interaction are observable; however, social intelligence, social competence and social skills are latent and can only be measured indirectly through external components.

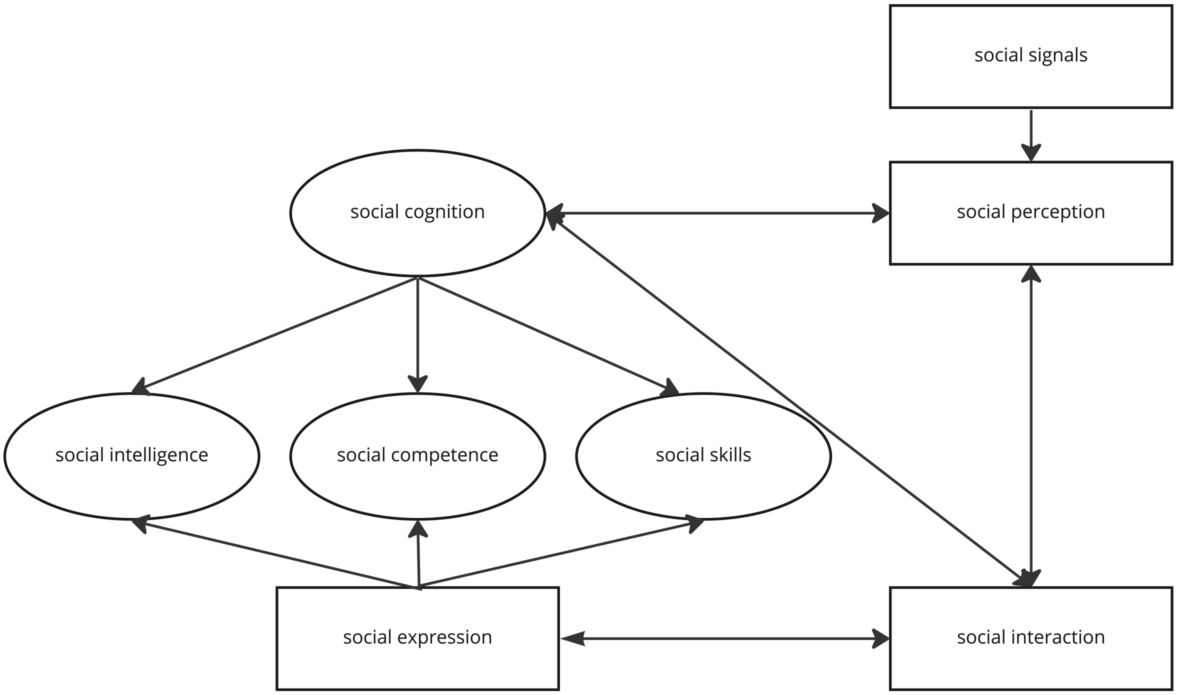

Figure 1: Draft of potential general framework for social cognition and concepts from previous weeks

1.3 Validation of Framework

We list the following few approaches to test the validity of above proposed definition framework:

Thought experiments. One initial attempt to validate a framework such as above would be to apply it to edge cases or hypothetical social interactions. While this does not quantitatively validate the framework, it may reveal certain flaws or gaps of a design that should be addressed

Scripted pipeline tests. Another way to validate these types of framework is through scripted hypothetical scenarios. This is a form of scenario-based design validation, that can emphasize how well a framework adapts to specific scenarios.

Diagram-based labeling of real-life social interactions. For a curated pool of practical social interaction experiences, a solidly reliable definitive framework for social cognition and interactions shall achieve full coverage in annotating every atomic procedural and static component, logical process, as well as internal and external interactions with various other constructs, with its proposed taxonomies. Simultaneously, it is expected for all proposed taxonomies to be frequently present in the majority of concrete events, in justification of necessity.

Challenge: Subconscious Constructs. As most subconscious constructs tend to be presented as latent variables in proposed models, which are also prone to noticeable individuality, they are challenging to describe, capture and hence measure. Additionally, disagreement between the groundtruth states and proposed ones of such constructs doesn't always translate into observable errors that directly challenge model validity, therefore requiring delicately devised tests to elicit conflicts that are tangible only when models are unreasonably defined.

2 Social Interactions and How to Operationalize in AI

In our discussion of Probe 3, we discuss the Social Interaction model designed by Ligthart et al. [2021]. This interaction model covers requirements to qualify an interaction as social and the basic components needed for the maintenance of social interactions. We propose additional aspects to this model to better operationalize this social interaction model:

Intention vs. Initial Interest. Intention extends beyond initial interest with inclusion of broader objectives to achieve. The latter is key to *invoke* interaction; the former contributes to *maintain* it. Still, it's possible to consider intention as an item of interest: partial interest is to achieve certain goals. While the Social Interaction model highlights interest, it does not yet capture intention.

Expected Agency of Artificial Social Agents in Social Interaction Definition of *agency* for artificial social agents in an interactive context shall incorporate human expectations of their functionality in certain scenarios. That is, agency as defined in the Social Interaction model does not take into account the human's expectation which may vary depending on the task or scenario.

3 Online and Offline Measurements

In creating a framework for social cognition, it is important to clarify how to operationalize and measure these constructs. For this discussion, we focused on two kinds of measurements: offline measurements and online measurements. In an online approach, evaluation is based on direct interaction with a social agent, often involving user-design study and analysis. This approach was taken in evaluating core elements of social interaction by Ligthart et al. [2021]. In contrast, offline approaches evaluate systems by sharing written scenarios directly with the social agent, e.g., SocialIQA [Sap et al., 2019].

With these types of evaluation in mind, we considered which parts of our framework in Figure 1 might be better evaluated online versus offline, as well as whether it is possible to evaluate these constructs offline or offline. We posit the offline measurements are particularly useful in evaluating third-person perspective social scenarios. For example, social interaction can by evaluated via constrained contexts and options. Similarly, social perception from a third-party view might be evaluated with battery of tests, similar to those commonly used in general intelligence [Conzelmann et al., 2013]. On the other hand, physical embodiment in social interactions might be impossible to evaluate using offline evaluation and may be better captured in a human-agent study. Thus, when measuring these constructs from a physical interaction standpoint, online measurements are necessary. Consequently, when selecting whether to evaluate a construct using offline or online techniques, it may be important the task and scenario that is being evaluated.

4 Modeling Individual Perspectives in Social Perception

To approach this probe, we use the framework provided by Brunswik [1956] as a foundation. In particular, the Brunswick model provides an empirical scaffolding for our thoughts through its core premise of probabilistic functionalism and its constructs, distal stimuli (objective states of an agents environment) and proximal stimuli (features of the environment perceivable to the agent). One general challenge of modeling social perception is the lack of ground truth labels. That is, social actors can only capture distal cues via perceivable proximal stimuli and consequently the ground truth cannot be directly measured.

We build off the Brunswick model to suggest the following graph-based modeling method for social perception which considers each social actor as nodes in a graph connected to multiple attribute nodes, which interact with further actor nodes to study similarity and to facilitate minute changes. In more detail, the process of sender expressing social signals and receivers perceiving them can be represented with:

$$S_{\text{perceived}} = G_{\text{receiver}}\Big(C_{\text{context}}, \text{ToM}(P_{\text{sender}}, P_{\text{receiver}})\Big)\Big(f_{\text{sender}}(S_{\text{conceived}}, P_{\text{sender}}, C_{\text{context}}) \cdot W_{\text{sender\_expression}}\Big) \cdot W_{\text{receiver\_perception}}$$

Specifically, the symbols denote the following concepts:

Variables	Meaning
$S_{\text{perceived}}$	the perceived social signals in receiver social agent
$S_{\text{conceived}}$	the conceived (yet not expressed) social signals in sender social agent
C_{context}	compound context variable that represents all information in the social scenario
P_{sender}, P_{receiver}	sender & receiver personality, impacting how Theory-of-Mind functions
$W_{\text{sender\_expression}}$	how the expressive habits or tendency of the social sender impacts the social cues
$W_{\text{receiver\_perception}}$	how the perception capacity of the social receivers impacts the raw social signals

Function	Objective
$\text{ToM}(\cdot)$	Theory of Mind processing, considering both sender and receiver personalities
$\text{G}_{\text{receiver}}(\cdot)$	perception determinant function, determining how the receiver perceives social information
$f_{\text{sender}}(\cdot)$	function to determine how it converts conceived social signals in receiver into social cues

Table 1: Symbols in Formulaic Representation of Social Interactions

5 Theory of Mind: Theory-Theory and Simulation-Theory

As outlined in Wiltshire et al. [2015], there are two general approaches to modeling theory of mind in another agent: Theory-Theory (TT) and Simulation-Theory (ST). **TT** [Gopnik and Wellman, 1992] is characterized by one agent understanding another's mental state through a combination of observation and reasoning based on some knowledge base about how the world works. For example, an agent may reach the understanding that someone is hungry because the other person is more distracted and irritable than normally, the hour is well after lunchtime, and the other person has yet to eat lunch. **ST** [Blakemore and Decety, 2001] is an alternative approach in which agents can arrive at an understanding of another's mental state by "putting themselves in the other person's shoes", and imagining how their mental state would change as a result. In the same example above, the agent could understand that their counterpart is hungry by imagining how they would feel if they had not yet eaten and it was well past lunchtime, then validating that against their observations (e.g. "would I be easily distracted and irritable if I was hungry?").

Our conversation largely centered around whether TT and ST can be unified into a single modeling paradigm, or if not, how we can distinguish whether TT and ST can co-occur in humans. A hybrid approach may involve using TT when the conditions you would have to imagine to simulate the experience differ so dramatically from your lived experience that any simulation would be inaccurate (e.g. if someone eats at the same time every day, the above example may be nearly impossible for them to simulate). Some relevant works in hybrid approaches are [Carruthers and Smith, 1996, Nichols and Stich, 2003]. Another perspective that was mentioned is that all TT is actually ST, because the "knowledge bases" we create are in a way simulations as well, of ourselves and of others. For example, the idea that "generally people are hungry long after lunchtime" is arrived at largely through lived experience and simulated models of other people's experiences). However, our conversation fell short in finding ways to test whether TT or ST is occurring in humans, and to what extent.

6 ToM and Artificial Social Intelligence

Can ToM Be Non-Social in Nature? We began by asking whether ToM can happen in a non-social way. One example is **education**: understanding how much someone knows about mathematics may not be an inherently social task, though some prominent mathematics educators argue that aspects of social intelligence can be critical in mathematics education [Wilf, 2010]. This may be as important in lecturing as it is in one-on-one communication, and there are many works that discuss the importance of gauging group dynamics in lectures [Lowe and Borkan, 2021]. However, moving from one-on-one Theory-of-Mind to group setting Theory-of-Mind may introduce new challenges. How does ToM scale to group settings? Is group ToM an aggregation of individual ToM's of the group members, or should group ToM be a fundamentally different problem? Works such as Woolley et al. [2015] would suggest that group ToM is fundamentally different. It may be difficult to distinguish social from non-social uses, modeling, and evaluation of ToM.

Is Interaction Critical to Understanding ToM? We also considered whether ToM skills can be evaluated offline, or whether online interaction – especially in social settings – is critical to modeling ToM. One point that was brought up is the following: we evaluate ToM in non-social settings in the following way: ask someone a question such that if they have some mental state s, they will answer one way and if they have some other mental state s' they will answer another way. For example, in mathematics, educators might give a question about multiplication on an exam such that if the student understands multiplication (s), they will answer it one way, and if they don't (s'), they will answer it another way. Currently self-reporting is the standard for understanding a person's state of mind. However, this could lead to biased or incorrect answers. Is it possible to put participants in situations that act as a "question", where it is very difficult to be incorrect about the "answer"? For example, in social interactions, if someone makes a joke, they can tell whether the other person genuinely finds it funny or not, and this tells them something about the person's mental state that the other person may not be able to hide. This could give them a better reading of the person's mental state than asking them "how are you feeling?" In this view, interaction would be critical to modeling ToM because it serves as a form of asking questions whose answers cannot be dissembled.

References

Simon Baron-Cohen. Theory of mind in normal development and autism, 2001.

Sarah-Jayne Blakemore and Jean Decety. From the perception of action to the understanding of intention. *Nature reviews neuroscience*, 2(8):561–567, 2001.

Egon Brunswik. *Perception and the representative design of psychological experiments.* Univ of California Press, 1956.

Peter Carruthers and Peter K Smith. *Theories of theories of mind.* Cambridge university press, 1996.

Kristin Conzelmann, Susanne Weis, and Heinz-Martin Süß. New findings about social intelligence. *Journal of Individual Differences*, 2013.

Hanne De Jaegher, Ezequiel Di Paolo, and Shaun Gallagher. Can social interaction constitute social cognition? *Trends in Cognitive Sciences*, 14(10):441–447, 2010. ISSN 1364-6613. doi: https://doi.org/10.1016/j.tics.2010.06.009. URL https://www.sciencedirect.com/science/article/pii/S1364661310001464.

Alison Gopnik and Henry M Wellman. Why the child's theory of mind really is a theory. 1992.

Michael F Green, David L Penn, Richard Bentall, William T Carpenter, Wolfgang Gaebel, Ruben C Gur, Ann M Kring, Sohee Park, Steven M Silverstein, and Robert Heinssen. Social cognition in schizophrenia: an nimh workshop on definitions, assessment, and research opportunities. *Schizophrenia bulletin*, 34(6): 1211–1220, 2008.

Mike EU Ligthart, Mark A Neerincx, and Koen V Hindriks. Core elements of social interaction for constructive human-robot interaction. *arXiv preprint arXiv:2110.04054*, 2021.

Robert C Lowe and Steven C Borkan. Effective medical lecturing: practice becomes theory: a narrative review. *Medical Science Educator*, 31:935–943, 2021.

Shaun Nichols and Stephen P Stich. Mindreading: An integrated account of pretence, self-awareness, and understanding other minds. 2003.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. Social IQa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4463–4473, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1454. URL https://aclanthology.org/D19-1454.

Sara M. Schaafsma, Donald W. Pfaff, Robert P. Spunt, and Ralph Adolphs. Deconstructing and reconstructing theory of mind. *Trends in Cognitive Sciences*, 19(2):65–72, 2015. ISSN 1364-6613. doi: https://doi.org/10.1016/j.tics.2014.11.007. URL https://www.sciencedirect.com/science/article/pii/S1364661314002502.

Herbert Wilf. On buckets and fires. 2010.

Jessica Williams, Stephen M Fiore, and Florian Jentsch. Supporting artificial social intelligence with theory of mind. *Frontiers in artificial intelligence*, 5:750763, 2022.

Travis J. Wiltshire, Emilio J. C. Lobato, Daniel S. McConnell, and Stephen M. Fiore. Prospects for direct social perception: a multi-theoretical integration to further the science of social cognition. *Frontiers in Human Neuroscience*, 8, 2015. ISSN 1662-5161. doi: 10.3389/fnhum.2014.01007. URL https://www.frontiersin.org/articles/10.3389/fnhum.2014.01007.

Travis J. Wiltshire, Samantha F. Warta, Daniel Barber, and Stephen M. Fiore. Enabling robotic social intelligence by engineering human social-cognitive mechanisms. *Cognitive Systems Research*, 43:190–207, 2017. ISSN 1389-0417. doi: https://doi.org/10.1016/j.cogsys.2016.09.005. URL https://www.sciencedirect.com/science/article/pii/S1389041716300493.

Anita Williams Woolley, Ishani Aggarwal, and Thomas W Malone. Collective intelligence and group performance. *Current Directions in Psychological Science*, 24(6):420–424, 2015.