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

Week 12: Relationships, Trust, and Ethics

Instructors: Louis-Philippe Morency and Paul Liang *Synopsis Leads: Ji Min Mun, Kelly Shi*

Edited by Paul Liang *Scribes: Gaoussou Youssouf Kebe, Ryan Liu*

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

Summary: Multimodal machine learning is the study of computer algorithms that learn and improve through the use and experience of multimodal data. It brings unique challenges for both computational and theoretical research given the heterogeneity of various data sources.

In week 12's discussion session, the class aimed to understand potential issues related to trust and ethics of artificial social agents. The following was a list of provided research probes:

1. During the past weeks, we studied technologies to enable artificial social agents. To help us understand the potential issues related to trust and ethics, it is good to first reflect on what applications are best suited for these technologies. What are the domains where you see artificial social agents having the most impact, in the short term and in the long term? In general, in which applications modeling is it particularly important to model social intelligence?
2. As socially-intelligent technologies are being deployed in houses and also in companies, people will start iterating with them multiple times, potentially over a long period of time, and maybe even creating relationships. What would be all the dimensions to consider when thinking of human-agent relationships? Should we replicate human-human interactions and relationships? How to define this relationship? A taxonomy of human-agent relationships?
3. How can we build social agents that are able to build long-term relationships? What are the main technologies missing to be able to build such relationships?
4. Another important facet is trust between humans and technology. What are all the facets of trust that are particularly important for social interactions with artificial social agents? Do they differ for human-human trust? Or differ from other AI technologies? How can we build trust between artificial social agents and people? What design principles should we follow to ensure trustworthy technologies? How to be careful not to deceive users and properly manage expectations?
5. What are the potential ethical issues related to artificial social agents? From all the ethical issues related to AI technologies in general, which ones are particularly important in artificial social intelligence?
6. What are the design principles and research strategies that we should consider when building artificial social intelligence? As academic researchers, how can we help design and implement these principles and strategies? In our own research, how to integrate these principles?

As background, students read the following papers:

1. Are friends electric? The benefits and risks of human-robot relationships [Prescott and Robillard, 2021]
2. Ethics of Artificial Intelligence and Robotics [Müller, 2021]
3. Formalizing Trust in Artificial Intelligence: Prerequisites, Causes and Goals of Human Trust in AI [Jacovi et al., 2020]
4. A Systematic Review of Attitudes, Anxiety, Acceptance, and Trust Towards Social Robots [Naneva et al., 2020]
5. Establishing and Maintaining Long-term Human-Computer Relationships [Bickmore and Picard, 2005]
6. An Introduction to Ethics in Robotics and AI [Bartneck et al., 2021]
7. Granny and the robots: ethical issues in robot care for the elderly [Sharkey and Sharkey, 2012]

8. How to Evaluate Trust in AI-Assisted Decision Making? A Survey of Empirical Methodologies [Vereschak et al., 2021]
9. Human vs. AI: Understanding the impact of anthropomorphism on consumer response to chatbots from the perspective of trust and relationship norms [Cheng et al., 2022]

We summarize several main takeaway messages from group discussions below:

1 Domains of Application of ASI (Probe 1)

The major domains of application for Artificial Social Intelligence (ASI) include perception and expression. In terms of perception, ASI can be used for tasks such as sentiment analysis, social media monitoring, facial expression recognition, tonality detection, intention detection, and mental state modeling. ASI is particularly effective in capturing subtle social signals and situations. ASI can also be used to infer a person's intentions if the prompt is unclear.

In terms of expression, ASI can be used for building social bonds, such as in entertainment, education, and healthcare. In the short term, it can be used for entertainment purposes with relatively few responsibility issues. In the long term, ASI could potentially be used in healthcare and education. However, the reliability of ASI needs to be improved before it can be used in these domains. ASI has the potential to create both dystopian and utopian scenarios, and it will be important to carefully consider the ethical implications of its use.

These ASI agents can then be used for providing companionship, especially when human capacities are limited and resource intensive, and providing adequate interactive services. More specifically, these agents would be most useful as home companions for household tasks, customer service agents such as tour guides that understand different cultural and social cues, mental health care agents in between sessions, and companions for relieving social isolation. In all these settings, these roles will be augmented by agents rather than be replaced by them. Therefore, the quality of service and care will increase without reducing the job market.

2 Human-Agent Relationships (Probe 2)

2.1 Taxonomy

The taxonomies of human-ASI (Artificial Social Intelligence) relationships can be broadly classified into two main categories: relationship design and relationship type.

In terms of relationship design, the focus is on how the human-ASI relationship is designed, taking into account various factors such as the type of interaction, the degree of autonomy of the ASI, and the nature of the tasks performed. One example of a taxonomy of human-ASI relationship design is the following:

Design Goal Types	Descriptios
Service-oriented	ASI designed to perform tasks for humans, such as scheduling, organizing, or providing information.
Social-oriented	ASI designed to interact with humans in a social context, such as for companionship, entertainment, or therapy.
Collaboration-oriented	ASI designed to work alongside humans in a collaborative manner, such as for creative projects, problem-solving, or decision-making.

Table 1: Human-Agent Relationship Design Taxonomy

In terms of relationship type, the focus is on the nature of the relationship between humans and ASI. One example of a taxonomy of human-ASI relationship type is the following:

Other taxonomies of human-ASI relationships may also consider additional factors, such as the degree of

Relationship Types	Description
Virtue friends	ASI designed to foster positive character traits in humans, such as empathy, kindness, or compassion.
Utility friends	ASI designed to provide a useful service to humans, such as for productivity or convenience..
Pleasure friends	ASI designed to provide entertainment or enjoyment to humans, such as for gaming, socializing, or leisure activities.

Table 2: Human-Agent Relationship Type Taxonomy

emotional engagement, the potential for loneliness reduction, or the risk of dependence on ASI. During the discussion, it was emphasized that ASI should not try to replicate all aspects of human behavior, especially negative traits like stubbornness. Instead, ASI should be designed to be compatible with humans and our flaws. However, there is a risk that this approach may create a hierarchy between humans and ASI, resulting in a unilateral relationship. Another addressed point is the need to balance reduction of loneliness and human detachment with the risk of creating overly docile robots. It was suggested that ASI is not the only way for loneliness reduction since it could also be achieved through encouraging people to interact with other humans.

Ultimately, the taxonomies of human-ASI relationships aim to provide a framework for understanding the different types of relationships that can be formed between humans and ASI, and how these relationships can be designed to optimize their benefits and minimize their drawbacks.

2.2 Human-Agent Collaboration

With the current limitations of ASI, it will be important for agents to know when to ask for humans to be in the loop. This will be extremely important for agents designed for high-stake situations in medical and healthcare domains. Most useful might be a hybrid of avatar-agent model where humans sometimes control the agents when necessary. In all the design goals mentioned in Table 1, there might arise situations in which users do not want to or cannot give full authority to agents to make the correct decisions. For example, for a social agent designed to alleviate loneliness, it will be crucial to reach out for human help when the user shows signs of self-harm or suicide. Therefore, it will be critical for humans and AI to develop a trusting cooperative relationship and for agents to learn effective communication with humans.

3 Long-term Relationships (Probe 3)

Long-term relationship between agents and humans can be fostered through continued interaction and personalization. One key challenge for ASI in achieving relationships with humans is to know when humans want to interact with them, as currently, users have to initiate the interaction. To address this challenge, information theoretic signals can be used to indicate the user's intent, but the level of signal required may vary. Another challenge for ASI is to handle interactions that fall outside their training domain. To solve this, ASI can be trained on a distribution of customer issues in fields like customer service to make them more effective in handling queries.

4 Trust (Probe 4)

Trust is a crucial factor in the adoption and use of Artificial Social Intelligence (ASI). Trust in ASI is subjective and depends on what individuals trust them on. Some people may not trust ASI due to privacy concerns as someone will have access to the data. Some people trust ASI due to their assumption that ASI is selfless.

The concept of trust in ASI can be classified into two categories: performance and disclosure. The former pertains to the dependability of AI in carrying out tasks, while the latter pertains to the extent to which individuals are willing to reveal their thoughts, feelings and confidential information to the technology. There

is also a concept of trust called Virtue friend, which is a form of trust where values and virtues match. This is difficult for ASI to achieve for now as they do not have a set of virtues. Moreover, the trust mechanism between human and ASI is fundamentally different from human-to-human trust. Humans trust each other, but robots trusting people is not bidirectional.

There are several crucial technical aspects we need to consider when building trustworthy ASI. Memory is crucial for building trust in ASI. If ASI keeps asking the same questions, it may not be considered a sincere agent. Knowledge retainment is crucial, and prompt prepending is not sufficient. If ASI is exposed to too much information, it may have seen too many opinions/identities, and there will be no consolidation of personality.

In conclusion, trust in ASI is complex and subjective, and it varies depending on individuals and their experiences. To build trust in ASI, transparency, sincerity, and knowledge retention are essential, but it is also important to recognize the fundamental differences in trust between humans and machines. ASI should also be designed to have a clear set of values and virtues that match human values and ethics.

5 Ethical Issues (Probe 5)

The development of artificial social intelligence (ASI) raises ethical concerns, including issues related to commercialization and monetization. One potential problem is the intentional design of addictive software for manipulative profit maximization. Another issue is virtual cloning, which involves training a model to duplicate a human's mind.

The treatment of ASI that exhibits social intelligence is a question. Should they be considered machines or given certain rights and protections? There are also concerns about how to dispose of or replace ASI if necessary.

The development of ASI may also change how we value things. Companies will have decision-making power over the behavior of these machines. In elder care, people may view it as less of a human responsibility to take care of the elderly if robots become more involved in this domain. Ethical issues must be proactively addressed to ensure responsible and ethical development and deployment of ASI.

6 Principles and Research Strategies (Probe 6)

When building artificial social intelligence, it's crucial to consider ethical principles and research strategies. One primary ethical concern is ensuring social agents do not manipulate users or violate their privacy or evade liability. To prevent manipulation, it's essential to have secure privacy, and take liability into account. Anthropomorphism is also an important consideration, particularly in education, where children may view AI as an authoritative figure without critical thinking skills. Therefore, increasing human involvement in agents may be necessary to ensure that children understand they are under parental control.

As academic researchers, we could work on developing new techniques for testing AI systems, including adversarial testing, to ensure that they are robust and effective. We can educate policymakers and the public about the importance of responsible AI development and advocate for policies that protect the public from potential harm.

In our own research, we can integrate these principles by prioritizing ethical considerations and designing experiments that test AI systems in realistic scenarios. We can also ensure that our data collection methods are respectful of users' privacy and that our AI models do not perpetuate biases or stereotypes.

References

Christoph Bartneck, Christoph Lütge, Alan Wagner, and Sean Welsh. *An Introduction to Ethics in Robotics and AI*. 01 2021. ISBN 978-3-030-51109-8. doi: 10.1007/978-3-030-51110-4.

Timothy W. Bickmore and Rosalind W. Picard. Establishing and maintaining long-term human-computer

relationships. *ACM Trans. Comput.-Hum. Interact.*, 12(2):293–327, jun 2005. ISSN 1073-0516. doi: 10.1145/1067860.1067867. URL https://doi.org/10.1145/1067860.1067867.

Xusen Cheng, Xiaoping Zhang, Jason Cohen, and Jian Mou. Human vs. ai: Understanding the impact of anthropomorphism on consumer response to chatbots from the perspective of trust and relationship norms. *Information Processing Management*, 59(3):102940, 2022. ISSN 0306-4573. doi: https://doi.org/10.1016/j.ipm.2022.102940. URL https://www.sciencedirect.com/science/article/pii/S0306457322000620.

Alon Jacovi, Ana Marasovic, Tim Miller, and Yoav Goldberg. Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI. *CoRR*, abs/2010.07487, 2020. URL https://arxiv.org/abs/2010.07487.

Vincent C. Müller. Ethics of Artificial Intelligence and Robotics. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Summer 2021 edition, 2021.

Stanislava Naneva, Marina Sarda Gou, Thomas L. Webb, and Tony J. Prescott. A systematic review of attitudes, anxiety, acceptance, and trust towards social robots. *International Journal of Social Robotics*, 12 (6):1179–1201, Dec 2020. ISSN 1875-4805. doi: 10.1007/s12369-020-00659-4. URL https://doi.org/10.1007/s12369-020-00659-4.

Tony J. Prescott and Julie M. Robillard. Are friends electric? the benefits and risks of human-robot relationships. *iScience*, 24(1):101993, 2021. ISSN 2589-0042. doi: https://doi.org/10.1016/j.isci.2020.101993. URL https://www.sciencedirect.com/science/article/pii/S2589004220311901.

Amanda Sharkey and Noel Sharkey. Granny and the robots: ethical issues in robot care for the elderly. *Ethics and Information Technology*, 14(1):27–40, Mar 2012. ISSN 1572-8439. doi: 10.1007/s10676-010-9234-6. URL https://doi.org/10.1007/s10676-010-9234-6.

Oleksandra Vereschak, Gilles Bailly, and Baptiste Caramiaux. How to evaluate trust in ai-assisted decision making? a survey of empirical methodologies. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW2), oct 2021. doi: 10.1145/3476068. URL https://doi.org/10.1145/3476068.