# Cooperative Social Intelligence: Understanding and Acting with Others

**Max Kleiman-Weiner (maxkw@mit.edu), Yibiao Zhao (ybz@mit.edu) & Joshua Tenenbaum (jbt@mit.edu)**
Brain and Cognitive Science,
Cambridge, MA 02139 USA

**Keywords:** multi-agent, cooperation, communication, coordination, theory-of-mind, social learning

## Theme

This workshop will focus on new developments and approaches to studying social intelligence with a specific focus on cooperation, theory-of-mind and social learning. With a diverse set of speakers and panelists, we anticipate these three themes will allow for connections to be made between developmental psychologists, cognitive scientists and artificial intelligence and robotics researchers.

1. **Cooperation:** How do we coordinate our limited individual capacities and perspectives to accomplish goals that no-one could have completed on their own? How do we share the spoils of a cooperative activity fairly and equitably and how does this capacity develop in early childhood? How do we learn who is cooperative and who is not to be trusted? What special purpose representations have we evolved that make cooperation so robust?

2. **Theory-of-Mind:** How do we go from sparse, noisy, underdetermined observations of behavior to acquire abstract knowledge of latent mental states which generalize to novel situations and people? How does this capacity develop and what representations support these capacities in infancy and early childhood? How do we understand the actions and intentions of groups or even collectives of agents?

3. **Social Learning:** When and how do we realize that other intelligent agents are often the richest source of world knowledge in an environment? How do we actively learn from others? How do we efficiently share important cultural knowledge through teaching? How do norms and conventions originally form and how do we learn existing norms and conventions so quickly?

One motivation for understanding social intelligence is to re-engineer socially intelligent artificial agents that treat people like people and can be treated like people by people. While a world where artificial agents roam the sidewalks still feels far away, automated agents are already roaming the streets in self-driving cars. The above challenges to understanding our own social intelligence become challenges for engineering cooperative AI:

1. **Cooperation:** How can we build agents that can work with us on mixed teams? Will they need to be taught cooperative values or must these values be baked in from the start?

2. **Theory-of-Mind:** How can we build agents that can understand our intentions, how they unfold over time, and the ways in which they may change dynamically? Can agents without theory-of-mind robustly cooperate with humans? How can agents reveal their intentions in ways that are natural to us?

3. **Social Learning:** How can machine learning from teachers go beyond imitation and reinforcement? Can artificial agents take advantage of the human ability to teach and learn like children do?

## Speakers

We have already invited and received confirmations from six speakers (one tentative) that will form the core of our workshop. These speakers come from communities ranging from computer science and artificial intelligence to cognitive science and developmental psychology. Each speaker brings a unique perspective that will be of interest to the entire CogSci community. We believe these interdisciplinary interactions are a unique and positive element of the CogSci community and we hope to build on that foundation. We may invite 1-2 more speakers and are committed to organizing a gender balanced workshop.

**Nick Chater (Professor of Behavioral Science, Warwick)** Possible topics: Virtual bargaining, instantaneous conventions, one-shot communication, joint intentions

**Joel Leibo and Thore Graepel (DeepMind)** Possible topics: Deep learning for cooperation and competition

**Stuart Russell (Professor of Computer Science, Berkeley)** Possible topics: Cooperative inverse reinforcement learning, value alignment, cooperative robotics

**Hyo Gweon (Professor of Psychology, Stanford)** Possible topics: Social learning, social development, theory-of-mind, cooperation

**Igor Mordatch (Professor of Robotics, CMU & OpenAI)** Possible topics: Emergent communication, deep reinforcement learning

**Victoria Southgate (Professor of Psychology, Copenhagen) [tentative]** Possible topics: Infant social cognition, action processing, Imitation and mimicry, Theory of mind, Motivation and learning

## Workshop Program

We plan to host a full day workshop consisting of talks given by 6-8 invited speakers (depending on final commitment). Our intention is to explore having an extended lunch and allow (space permitted) for the presentations of posters in this topic area. This will allow for more informal interactions between interesting parties and will allow for researchers to have an additional opportunity to present their work. At the end of the workshop we will have a discussion panel with all of the speakers to synthesize the topics discussed throughout the workshop.

## Potential Financial Support

As this topic will be of interest to some of the industrial research labs we will also see if they are interested in providing some support to defer some of the registration costs. However as we already have confirmations from most of our speakers this funding will not be necessary and we believe we will have a very successful within the constraints of the funding provided by the conference.