Dissertation Summary

This thesis consists of three empirical chapters that investigate elements of human social behavior through the combination of economic games, computational modeling, and neuroimaging. **Chapter 2** uses the attacker-defender contest and a cognitive-hierarchies framework (Camerer, Ho, & Chong, 2004). The cognitive-hierarchies framework quantifies the depth of mentalizing recursion, i.e., I think that you think that I think. We found that during attack relative to defense individuals invested less and were less successful, and that investments in attack utilized more levels of cognitive recursion (i.e. more sophisticated mentalizing) than investments in defense. Furthermore, attack behavior was preferentially associated with neural activity in the ventral striatum, a region consistently linked with reward learning, and the temporoparietal junction, a region consistently linked with perspective-taking and social cognition. We conclude that in economic contests, coming out ahead (versus not falling behind) involves sophisticated strategic reasoning that engages neural regions associated with both value computation and theory of mind.

A key task for defenders in the attacker-defender game studied in Chapter 2 is to assess to what extent they can trust their counterpart to not attack, or should instead fear their counterpart’s aggressiveness. **Chapter 3** zooms in on trust and distrust as a key element in social interactions. We show that variability in reciprocity (participants playing as responders) can be exhaustively captured by three categories: exploiters (individuals who never reciprocate), perfect reciprocators (individuals who always reciprocate), and contingent reciprocators (individuals who reciprocate as a function of how much they are trusted). This variability is learned by senders through a combination of reinforcement and belief-based learning. However, senders learn to trust imperfectly, frequently failing to arrive at the optimal policy, in particular when interacting with contingent reciprocators. Furthermore, the degree to which individuals weigh belief over reinforcement is positively correlated to their average payoff, indicating that learning to trust from mentally simulated outcomes outperforms learning from observation only.

The results from both Chapter 2 and 3 revealed an important role for social perception and learning, suggesting that empathy and social norms modulate decisions to exploit and to trust and reciprocate. **Chapter 4** builds on these and related findings by asking what role empathy (Zaki, 2014; Zaki & Mitchell, 2013) and social preferences such as concerns for fairness and the welfare of others (Blake et al., 2015; Fehr & Schmidt, 1999) play in learning group-specific conventions. We created three populations with different rules of engagement and varied whether or not decisions affected interaction partner outcomes. Participants made ultimatum bargaining offers to responders from these different populations and could observe whether their offer was accepted or rejected. Participants quickly adapted to group-specific rules in learning environments without social consequences, but were overly generous and ended up misrepresenting what would be acceptable when decisions affected their partner’s outcomes. We propose a computational model, combining Bayesian principles and social preferences, that mechanistically explains how generosity leads to biased sampling, impeded learning, and false beliefs about what offers are deemed acceptable. Using functional neuroimaging, we mapped key computational variables in two major brain networks, previously associated with value-based and social decision-making. Results suggest that generosity, related to brain regions associated with decision-conflict and perspective-taking, can induce self-fulfilling beliefs in pro-sociality norms that may help to increase cooperation and reduce conflict between distinct groups but also create inaccurate stereotypes and economic inefficiencies.