Formalizing and Testing Computational Cognitive Models of Social Collaboration

By

Vael Gates

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Neuroscience

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Thomas L. Griffiths, Co-chair
Professor Anne G. E. Collins, Co-chair
Professor Ming Hsu
Professor Anca D. Dragan

Spring 2021

Abstract

Formalizing and Testing Computational Cognitive Models of Social Collaboration

by

Vael Gates

Doctor of Philosophy in Neuroscience

University of California, Berkeley

Professor Thomas L. Griffiths, Co-chair

Professor Anne G. E. Collins, Co-chair

The greatest human achievements are never completed alone. However, social interaction is complex and successful collaboration even more so. Tremendous amounts of information and processing are involved in predicting, interpreting, and working with others. These calculations are implicit, deeply embedded in the human psyche and not easily accessible for analysis or improvement. Knowing what algorithms the mind is solving when collaborating with others would be invaluable, both for our own knowledge and improvement of social collaboration between people, and to reconstruct these algorithms in systems outside ourselves. People increasingly interact with artificial intelligence systems, and developing models of social interaction could enable these systems to seamlessly assist us. This premise motivates my work: that by understanding the algorithms behind social collaboration, we can improve interactions between people, and between people and machines.

In this dissertation, I formalize and test computational models of collaboration, focusing on three problem areas. First, I investigate how people collaborate in recalling information. Though one might expect memory to be a solitary venture, researchers have long studied the differences in how people recall information in groups compared to alone, though only for small group sizes. Luhmann and Rajaram (2015) hypothesized mechanisms of large-scale recall using an agent-based model. I test the predictions of this model with an empirical experiment recruiting thousands of participants. Second, I investigate another key component of collaboration: people's intuitive judgments of how shared resources should be allocated among people with different preferences. I collect empirical data from participants under a number of decision conditions, then use an inverse reinforcement learning model to determine what underlying mathematical fairness principles characterize people's choices. Third, I investigate how people infer others' preferences, a key component of collaboration. My coauthors and I present a rational model for inferring preferences from response times, using a Drift Diffusion Model to characterize how preferences influence response time and Bayesian inference to invert this relationship. We then compare the model's predictions to collected empirical data. These three case studies comprise novel models of social interaction, tested with behavioral experiments, aimed at improving human and technological collaboration.

Dedicated to all the people who made this PhD possible. (And LaTeX, which I appreciate a lot and also is finicky enough that including this dedication page is definitely the right move for ensuring the page numbering works out.)

Contents

Acknowledgments

So many people have taken me under their wing during this PhD. I certainly did veer around a bit, from neuroscience to the "but should I quit?" period, to clinical psychology and back again to artificial intelligence (AI). Throughout it all, the person I'm most grateful to is my advisor Tom Griffiths. I remember the first time I had a conversation with him, sitting awkwardly across the desk from each other in silence, as he decided that I needed more human-robot interaction in my life (wait, what?) and set me up to interview with a collaborator who I ended up working with throughout my entire PhD. Tom has these tendencies—to steer, guide, and teach—and is also the most supportive person I could imagine for when I steered myself down only vaguely-related tributaries.

"Tom, I want to work on AI value alignment," "Tom, I want to be remote advised by you when you move to Princeton," "Tom, I'm thinking about taking a leave of absence," "Tom, I'm changing my name," "Tom, I want to do this program in clinical psychology," "Tom, I'm applying for social scientist jobs in AI," "Tom, you've yet again led me to insights about how diverse areas of the cognition, computation, and human applications interact, thank you," "*Tommmmm*." All these years, Tom's been an unwavering stalwart of encouragement, insight, wisdom, and freedom: guiding, seeing, and making and letting things happen for me. Sending me ":-)" (with noses!) when I was intimidated. Exemplifying capability and calm, and excellent teaching and mentoring. Demonstrating astonishingly flexible, far-sighted, and grounded thinking and research. I'll miss Tom a lot when I graduate. "Best advisor," I tell people and have been emphatically telling people for years, and I will forever value his mentorship during this period.

As I've flowed down the to-various-degree-related tributaries of my PhD, I've met so many other wonderful mentors and people. Anca Dragan, the PhD-long collaborator, cheerfully welcomed me into her lab, spent hours working through math with me on her whiteboard, and has helped me in countless small ways over the years. Anne Collins, who along with being a rotation advisor, provided me desk space during the year Tom moved to Princeton, served on my thesis committee, wrestled more than once with the administration for me, hired me on to teach, and has always been unfailingly and conscientiously supportive. Ming Hsu, who so warmly welcomed me to Berkeley and then went on to serve on my thesis committee for many more years. To my excellent thesis committee: thank you.

I've had several forks in my PhD life, but I'd say the largest was when I asked to take a leave of absence because I wasn't sure of the PhD. Tom told me this was completely normal, and that if I wanted to I absolutely could, but let's talk about how to make your PhD more fulfilling before you go. It turns out that I wanted to explore clinical psychology. After interviewing several professors and writing a public-facing article about clinical psychology research, I got to spend an entire year doing clinical psychology training with the ever-so-generous Clinical Psychology PhD program. I have so many thank yous for the people there who gracefully took me in, guided me, and worked alongside me. Sheri Johnson, Nancy Liu, Diana Partovi, Bruce L. Smith, Qing Zhou, Kealey McKown, Cindi Baker-Smith, Matt Elliott, Garret Zieve, Sinclaire O'Grady, Catherine Callaway, and Emily Rosenthal, among others: I had a truly wonderful time. Thank you especially to Jackie Persons, who took me on to do clinical psychology research and advocated for me on many fronts. (An additional thanks to the Berkeley Science Review, who helped me write that article that introduced me to the field.)

Throughout my PhD, I also maintained an interest in AI safety, and want to thank all the people at the Center for Human-Compatible AI who opened that space to me (including its overlap with the non-university-associated outside communities where I have found a home). Thank you especially to Andrew Critch, who's been giving me little pushes and ebullient guidance over many years, some of which have been game-changing. Thank you to Cody Wild, Steven Wang, Mark Nitzberg, Rosie Campbell, Tom Gilbert, Stuart Russell, Rohin Shah, and many others. I give all the thanks to Jaime Fisac, who was a never-ending fount of positivity and enthusiasm as well as mathematical models: I greatly enjoyed working with you for 3+ years. Thank you to Brian Christian, who talked with me about career plans and was similarly an inspiration and model (and the many other people who patiently answered my questions about careers).

Thank you to my other computational cognitive scientist research collaborators! Jordan Suchow, who led me through my first first-author research project and whose napkin on which he stamped "SCIENCE" I still have, and Fred Callaway, who entered Tom's lab in the same year I did and has been an enduring and friendly presence. Mark Ho, who was always happy to discuss research interests and help out, and the other senior students in Tom's lab with whom who I was lucky enough to do research (Aida Nematzadeh, Jess Hamrick) or spend time (Ellie Kon, Alex Paxton, Thomas Langlois, Rachel Jansen, Rachit Dubey, Falk Lieder). Thanks to Mike Pacer and Jordan Suchow for creating Dissertate, which I used to create this dissertation; to Jess Hamrick for creating nbgrader, and David Bourgin for his extensive and generous help in how to use it. Thanks to the Dallinger team, especially Jesse Snyder, for helping me debug experiments over the years. Thanks to Samee Ibraheem for the years-long mafia game work, and to my research collaborators from other institutions: Tess Veuthey, MH Tessler, Tobi Gerstenberg, Kevin Smith, Laurie Bayet, and Josh Tenenbaum, and to the MBL program that let us all meet.

I want to thank the Neuroscience Department at UC Berkeley, who were ridiculously kind to me. To Candace Groskruetz: *thank you*, this PhD would have been so painful without you. To Michael Silver, who made things work for me to a fault, and to Anne Collins, Marla Feller, Robin Ball, and Steve Brohawn, Nate Munet, Karina Bistrong, Logan Thomas, Neha Wadia, Rachit Dubey, and Maria Eckstein for teaming through teaching. To Frédéric Theunissen, who so excellently served on my qualification exam committee. To my quirky and wonderful cohortmates: Christine Tseng, Carson McNeil, Holly Gildea, Zuzanna Balewski, Tobias Schmid, Kevin Yu, Bill Croughan, Jacob Miller, and (adopted) Toby Turney: I very much enjoyed my time spent with you. To the many other members of the neuroscience program, including long-graduated mentors like Yvonne Fonken and Maureen Turner. To the NIH Training Grant, the Neuroscience Program, and other sources of funding.

Finally, thank you to all of the people supporting me outside of this PhD, before and during (including those names who I have forgotten to include: there are many!) Thanks to all my friends: this felt like a period of growth and happiness for me, and so many of you contributed. None of this could have happened without all my previous research mentors— Bevil Conway, Ellen Hildreth, and Zoe Kourtzi— and all of the mentors before them. And I'll end with a final thanks to Tom, who made this PhD the best PhD I could have done, given where I was and who I am now.

Thank you all so much. I look forward to the research ahead!

Chapter 0

Introduction

We live in a complex world, which we navigate with the help of others. However, social interaction itself is complex and often implicit. How do we think differently in the presence of others? How do we navigate situations where people's preferences differ? How can we infer what others think? These questions are some of many that people intuitively understand and act upon as they cooperate with each other. Yet as as we move into an age where people increasingly cooperate not only among themselves, but with artificial intelligence systems (e.g., Chandrasekaran & Conrad, 2015; Fong et al., 2003; Severinson-Eklundh et al., 2003; Wang et al., 2020; Wilson & Daugherty, 2018), having an explicit understanding of what "collaboration" means to people becomes increasingly useful. Specifically, having an explicit computational description of the tasks and algorithms we use when we collaborate with each other can help us improve our teamwork: between humans, and when humans are collaborating with artificial intelligence systems. These dual goals—improving collaboration between humans, and improving collaboration between humans and machines—motivate the work that follows, which manifests the drive to deeply and explicitly understand the algorithms behind social collaboration.

In this dissertation, I focus on three specific problem settings in the large space of necessary components of successful collaboration. For each of these areas, I formalize or test computational cognitive models of social interaction. I approach this work from the lens of computational cognitive science (Chater et al., 2006; Gershman et al., 2015; Griffiths, 2015; Griffiths et al., 2010; Griffiths et al., 2008; Perfors et al., 2011; Tenenbaum et al., 2006; Tenenbaum et al., 2011). More specifically, I combine probabilistic models with empirical data from crowdsourced experiments. This work builds on previous research using Bayes' rule to invert decision-making models (Baker et al., 2017; Baker et al., 2009; Jara-Ettinger et al., 2016; Jern et al., 2017; Kiley Hamlin et al., 2013; Lucas et al., 2014). It is highly interdisciplinary, spanning areas of collaborative memory, artificial intelligence, robotics, and behavioral economics as well as social cognitive science. In each of my problem settings, I aim to elucidate the algorithms people use to collaborate, so that we may improve collaboration between people, and gain the ability to embed these mathematical descriptions in artificial systems that cooperate with people seamlessly.

I focus on the following three questions, formalizing and testing computational cognitive models of social collaboration within these problem settings. I ask:

- How does collaborating with others modify how we recall information?
- How do we decide what is best when people have different preferences?
- How do we make inferences about others' preferences from their actions?

In Chapter 1, I investigate the first question, aiming to understand how people remember and interact with information differently when they are collaborating or alone. Researchers have found that people's ability to recall words differs if they are in small groups versus by themselves, but researchers had not investigated this effect for large groups due to logistical difficulties. I examine collaborative memory in large groups using a large-scale online study, and compare my empirical results to a computational model that had been hypothesized to describe both small and large group behavior. In Chapter 2, I investigate another key component of social collaboration, modeling our intuitive senses of what to do when trying to satisfy people with different preferences. People are often in environments where they can provide a resource for other people, but the recipients each have different preferences over the resource options: for example, a teacher choosing between field trips that different children would enjoy, or a government choosing between aid programs that would affect different citizens. In this work, I quantitatively describe people's intuitions for what should be done. This information can be applied to current-day collaborative environments, and also has applications if these decisions become automated in the future (e.g. self-driving cars choosing destinations for a family). Finally, in Chapter 3, I address how we make inferences about others' preferences from their actions, a required element in collaborating with others. People can know much more about another's preferences than what is said: for example, if you ask someone on a date, and they say "Yes!" instantaneously compared to "...yes!", you can infer something about their preferences from that response time. I and my coauthors create a rational model describing these inferences about other people's preferences from their response times, and collect empirical data to test the model's predictions.

I approached each of these problem areas with computational cognitive science methodology: starting from or developing a computational model of a cognitive phenomenon, and testing that model's predictions with empirical data. Computational models are very comprehensible, succinct representations of cognitive processes that condense our understanding of intuitive mechanisms. Just as importantly, computational models can be unambiguously tested, if novel human experiments are developed and completed to test their predictions. The use of inverted generative models based on Bayesian inference, combined with behavioral experiments, has been promising for describing our implicit cognitive algorithms, and combining insights into more comprehensive models. Having these techniques let us address the questions above with new lenses and insight, and the experiments we developed could not have been completed at all with these tools and framework.

Consider the content described in Chapter 1, "Memory transmission in small groups and large networks: an empirical study." Collaborative memory had been extensively studied in small groups (Rajaram & Pereira-Pasarin, 2010), but it was not practically feasible to study people's recall in large groups. A model was developed to formalize the mechanisms hypothesized to be underlying collaborative memory (Luhmann & Rajaram, 2015), which gave concrete predictions for what would happen in large groups. By testing that model and showing that the empirical results did not match those predicted, we demonstrated that researchers do not yet have a complete understanding of the mechanisms of collaborative memory. In other words, the model was a concrete and succinct summary of hypothesized mechanisms, an empirical study testing its predictions showed it did not completely capture reality, and this mismatch provides important evidence to be incorporated in the next, more accurate set of hypotheses we develop for the cognitive mechanisms underlying collaborative memory. Even without delving into the impressiveness and advantages of crowdsourced large-scale studies (enabled by the intersection of cognitive science and computer science), the process of formulating and testing a computational cognitive model of social interaction was very useful as applied to testing our understanding of collaborative memory.

Taking a computational cognitive approach was also key to answering the question in Chapter 2, "How do we decide what is best when people have different preferences?" This work drew upon varied previous research, most relevantly previous empirical studies of people's preferences concerning fair allocation, operationalized using mathematical formalizations of fairness (Engelmann & Strobel, 2004; Fehr & Schmidt, 2006; Herreiner & Puppe, 2007; Yaari & Bar-Hillel, 1984), and the framework of "inverse reinforcement learning" from computer science (Ng & Russell, 2000), which allows inference of goals from actions. We connected these threads in this work. Our empirical study let us probe what people thought was correct behavior for an uninvested third party, which we determined not by asking participants to indirectly self-report, but by observing what choices people selected for a third party agent in a nuanced choice environment. Formalized fairness metrics represented our quantitative, concrete hypotheses for what algorithms could be driving people's behavior. Then, an inverse reinforcement learning model let us extract those cognitive algorithms from people's actions. These components, key to computational cognitive science methodology, were needed to answer our question: what are the implicit algorithms driving what people think is correct behavior in serving diverse others?

Finally, the question in Chapter 3, "How do we make inferences about others' preferences from their actions?" suits the computational cognitive approach particularly well. Social inferences are notoriously complicated and hard to describe, but computational models allow a compact representation that is easy to understand and generates concrete predictions. There is a long history of using computational cognitive models to describe aspects of pragmatics and evaluating the accuracy of these models with empirical studies (e.g., Frank & Goodman, 2012; Goodman & Frank, 2016). Though our study did not involve language, testing and modeling inference from pauses in decision-making comes from the same lineage. We also drew from previous work studying the relationship between preferences and response time specifically: see drift-diffusion models, and e.g. Konovalov and Krajbich (2020). Thus, drawing from this background, we developed a computational cognitive model to describe how people could infer others' preferences from their response times, and tested our model's predictions with an empirical study. This work, with its inherent explanation of underlying cognitive algorithms and testing of predictions, could not have been conducted under a different methodological paradigm. Another benefit of this approach is that computational models can be cumulative, easily integrating with and building upon previous additions. In studying theory of mind inferences from response times, we add another verified description of the mechanisms of social inference to our collective knowledge. Ultimately, the aim is to unify these descriptions of specific phenomenon into a quantitative, complex, testable, and comprehensive model of social interaction.

In summary, in this dissertation I formalize and test computational cognitive models of social collaboration, focusing on three problem areas. The first problem area focuses on the question "How does collaborating with others modify how we recall information?" In navigating the world, people have the choice to collaborate or work alone. These decisions are highly dependent on the benefits and costs of collaboration (e.g., Colman, 1995, 2003), and this study investigates just how beneficial or costly it is to work together: what the outputs are, against the background of studying what mechanisms drive recall and how recall is affected by collaboration. The second problem area focuses on the question “How do we decide what is best when people have different preferences?” For almost any group decision we find ourselves a part of, people will have different preferences. Understanding what is best to do in that context is an essential human question, and one that strongly influences whether people will choose to be part of the group and collaborate at all. This study investigates what choices people think an uninvested third party should make when they are responsible for helping a group. The third problem area focuses on the question “How do we make inferences about others’ preferences from their actions?” Collaborating with others is always something of a mine(mind)field: others’ true preferences, goals, and beliefs are never directly observed, but are instead inferred via indirect actions like statements, decisions, and actions. Much of the art of social interaction and collaboration is thus about making backward inferences from actions (even actions as minor as “time before response”) to people’s more explanatory and fundamental values and beliefs (e.g., Baker et al., 2017). This study describes and tests a model of how people make inferences about others’ preferences from their response times. Cumulatively, these three problem areas use computational models to better formalize how people think and cooperate in a social world. I describe this work in Chapters 1, 2, and 3, then conclude.

Chapter 1

Memory transmission in small groups and large networks: an empirical study

When people try to remember information in a group, they often recall less than if they were recalling alone. This finding is called *collaborative inhibition*, and has been studied primarily in small groups because of the difficulty of bringing large groups into the laboratory. To study the dynamics of collaborative inhibition in large groups Luhmann and Rajaram (2015) constructed an agent-based model that extrapolated from previous laboratory experiments with small groups. The model predicts that collaborative inhibition should increase with group size. Here, we evaluate this model by recruiting a large number of participants using crowdsourcing, allowing us to replace the artificial agents in the model with people to study collaborative memory at larger scales.

It is perhaps surprising that people would process information differently if they are collaborating with others versus recalling words by themselves: one would think something as core as memory would not be impacted by others' presences. However, the collaborative inhibition effect is frequently observed (Rajaram & Pereira-Pasarin, 2010), and the agent-based model from Luhmann and Rajaram (2015) provides an intriguing hypothesis describing the mechanisms of how collaboration affects how people think and act. We were interested in empirically testing whether this model's predictions held true. We found that our empirical results did not match the model predictions: there was not evidence for an increase in collaborative inhibition with group size, as was predicted by the model.

We find these results motivating, both because it impresses the importance of testing hypothesized models of cognition, and because it highlights the opportunity to develop models that can be used to explain and eventually improve human cognition. If collaborative memory is consistently worse than solo recall, then we can build that insight into society, and even use technology to help engineer social environments to that end. In this chapter, I investigate the mechanisms of how collaboration affects recall, one of the many important components of social processing to understand in improving collaboration.

The contents of this chapter are as yet unpublished; preliminary results from Experiment 1 were presented at the Annual Conference of the Cognitive Science Society and appear in the Proceedings of that conference (M. A. Gates et al., 2017). The referenced "we" includes my coauthors Jordan Suchow (Stevens Institute of Technology) and Thomas L. Griffiths (Princeton University). This work was funded in part by NSF grant 1456709 to T.L.G., DARPA Cooperative Agreement D17AC00004 to T.L.G and J.W.S., and N.I.H. U.C. Berkeley Neuroscience Training Program Grant to V.G.

1.1 This work is embargoed until publication, but see M. A. Gates et al. (2017) for preliminary results.

Chapter 2

How to be helpful to multiple people at once

When someone hosts a party, when governments choose an aid program, or when assistive robots decide what meal to serve to a family, decision-makers must determine how to help even when their recipients have very different preferences. Which combination of people's desires should a decision-maker serve? To provide a potential answer, we turned to psychology: what do *people* think is best when multiple people have different utilities over options? We developed a quantitative model of what people consider desirable behavior, characterizing participants' preferences by inferring which combination of "metrics" (*maximax*, *maxsum*, *maximin*, or *inequality aversion (IA)*) best explained participants' decisions in a drink-choosing task. We found that participants' behavior was best described by the *maximin* metric, describing the desire to maximize the happiness of the worst-off person, though participant behavior was also consistent with maximizing group utility (the *maxsum* metric) and the *IA* metric to a lesser extent. Participant behavior was consistent across variation in the agents involved, and tended to become more *maxsum*-oriented when participants were told they were players in the task (Expt. 1). In later experiments, participants maintained *maximin* behavior across multi-step tasks rather than short-sightedly focusing on the individual steps therein (Expt. 2, Expt. 3).

This problem area addresses the question "How do we decide what is best when people have different preferences?" By repeatedly asking participants what choices they would hope for in an optimal, just decision-maker, and carefully disambiguating which quantitative metrics describe these nuanced choices, we help constrain the space of what behavior we desire in leaders, artificial intelligence systems helping decision-makers, and the assistive robots and decision-makers of the future. I consider our results important for the particular question we answered: distributing resources is both a common and often-contentious collaborative activity, and thus developing explicit models of what people prefer could be key to improving these interactions. However, I consider this work even more important as an example of a question class, in which human preferences are quantitatively queried and described. Humans are incredibly complex, and have different intuitions about what should be done or what is fair in many domains. Querying humans, and making quantitative models that describe their preferences, seem like potentially critical steps for developing artificial intelligence systems that are aligned with human values.

The contents of this chapter are published in the journal *Cognitive Science* (V. Gates et al., 2020). The referenced "we" includes my coauthors Thomas L. Griffiths (Princeton University) and Anca D. Dragan (University of California, Berkeley). Special thanks to Professor Anant Sahai for suggesting the conditions in Experiment 1, and we thank various members of the Center for Human-Compatible Artificial Intelligence for helpful comments. This work was funded in part by NSF grant 1456709 to T.L.G. and N.I.H. U.C. Berkeley Neuroscience Training Program Grant to V.G.

2.1 The full version of this work is available in V. Gates et al. (2020).

Chapter 3

A rational model of people's inferences about others' preferences based on response times

There's a difference between someone instantaneously saying "Yes!" when you ask them on a date compared to "...yes." Psychologists and economists have long studied how people can infer preferences from others' choices. However, these models have tended to focus on what people choose and not how long it takes them to make a choice. We present a rational model for inferring preferences from response times, using a Drift Diffusion Model to characterize how preferences influence response time and Bayesian inference to invert this relationship. We test our model's predictions for three experimental questions. Matching model predictions, participants inferred that a decision-maker preferred a chosen item more if the decision-maker spent less time deliberating (Experiment 1), participants predicted a decision-maker's choice in a novel comparison based on inferring the decision-maker's relative preferences from previous response times and choices (Experiment 2), and participants could incorporate information about a decision-maker's mental state of cautious or careless (Experiments 3, 4A, and 4B).

This problem area delves into a fascinating area of collaboration: our internal models of what other people are thinking, and how we infer this information from what they say or do. Specifically, we investigated preference inference from response times, making a quantitative model to describe this intuitive, commonplace phenomenon. This type of inference is so essential to collaboration that certain types of communication require it: inferring preferences from pauses is both assumed and required for some forms of veiled communication and humor. These types of inference models are exciting both in that they can be described with relatively simple math, and also in their potential applications to artificial agents. This work thus describes a final example of formalizing and testing a computational cognitive model key to social collaboration.

The contents of this chapter are as yet unpublished. The referenced "we" includes my coauthors Frederick Callaway (Princeton University), Mark K. Ho (Princeton University), and Thomas L. Griffiths (Princeton University). This work was funded in part by John Templeton Foundation [grant number 61454] to T.L.G.

3.1 This work is embargoed until publication, but see https://doi.org/10.31234/osf.io/25zfx for the preprint.

Chapter 4

Conclusion

We live in a complex world, which we navigate with the help of others. However, our social worlds are also complex: from a young age, we learn to develop models of what other people think and want, how they make decisions, and how they will cooperate or compete. By the time we reach adulthood, we can fluidly and intuitively use these social models to collaborate with each other. As society moves into the digital age, we moreover use and expect these models to be present and functioning in our artificial intelligence systems.

When something goes wrong in our communication with each other, or one of our devices fails to grasp a subtle social cue, people often find that making their social knowledge explicit is a difficult task. However, computational cognitive science has developed tools to make this implicit knowledge concrete. Formalizing mathematical models and testing them with empirical data has given and continues to give us a better understanding of how human minds interact with each other. Furthermore, having these quantitative formulations of how people cooperate has applications not only for improving human-human cooperation, but also for improving the algorithms in artificial intelligence systems, so that they are more able to effortlessly communicate and collaborate with people.

In this dissertation, I formalized and tested computational cognitive models of social collaboration, focusing on three problem areas. These problem areas are diverse, but are each important to understanding and improving collaboration.

In Chapter 1, I investigated how people collaborate in recalling information. Psychological studies have shown that people recall less information together compared to when they are by themselves, and Luhmann and Rajaram (2015) developed an agent-based model describing the mechanisms of how people remember words in groups and alone. I tested this computational model by conducting a large-scale behavioral experiment, and comparing my results to the model's predictions. I and my coauthors found that the empirical results and model predictions were not in alignment, inspiring future work to develop more accurate models of how social collaboration can affect memory. The question this work points to is broadly important: How does collaborating with others modify how we process information? Memory is a context where we might not expect collaboration to come into play, but developing models of how social interaction affect functions as core as memory could result in insights that would affect many.

In Chapter 2, I investigated another facet of collaboration: determining people's intuitive judgments of how shared resources should be allocated among people. In this project, I and my coauthors developed a computational model of what people think should be done, in terms of how resources should be allocated to diverse others. Participants completed multiple allocation tasks, and we used an inverse reinforcement learning model to characterize their decisions in terms of four fairness metrics. This work addressed the broad question of "How do we decide what is best when people have different preferences?" The ubiquity of this question drives its importance, and the more general approach of querying and developing quantitative models of what humans think is good seems key to developing understandable and collaborative systems, both human and artificial.

In Chapter 3, I investigated how people infer others' preferences from their actions, one of the most intriguing and intuitive aspects of collaboration. This work drew from an intuitive example: when someone instantaneously says "Yes!" when you ask them on a date compared to "...yes," even though the answer is the same, we can infer a lot about this person's preferences from their response time. Here, my coauthors and I presented a rational model for inferring preferences from response times, using a Drift Diffusion Model to characterize how preferences influence response time and Bayesian inference to invert this relationship, and tested the model's predictions in three experiments. The sheer intuitiveness and ease with which people execute inferences from pauses indicates the pervasiveness of such cues in social interaction... which in turn indicates that being able to quantitatively describe and implement inferences of other people's mental states, beliefs, and preferences could be essential to developing collaborative artificial intelligence agents.

This work also has limitations. Each specific project has methodological limitations and limits to what can be inferred from the results, which are discussed within the individual chapters. When viewing the overall research scope, however, the most striking limitation (and opportunity for future directions) is the inadequacy of addressing only three questions compared to the full breadth of the topic. While I presented three problem areas that fit within the framework of computational cognitive models of social inference, there are innumerable others. Many of these unanswered questions are well-suited to computational cognitive science techniques and will fall neatly within the framework. Other questions on social collaboration will call for more qualitative or observational work, since computational models are often very good at describing, explaining, or unifying robust phenomena but can shine less in the discovery stage (though on the other hand, they can offer powerful predictions and hypotheses).

Diving in a little more, we are still left wanting with respect to the three overall questions we did ask: "How does collaborating with others modify how we recall information?", "How do we decide what is best when people have different preferences?", and "How do we make inferences about others' preferences from their actions?" Beyond the work described in Chapter 1, asking "How does collaborating with others modify how we recall information?" immediately prompts the question of how recall changes with factors other than group size. The established collaborative memory paradigm has explored a number of variables, including the content of the presented information (presenting words of similar or different categories), the presented information format (presenting words versus images), timing and retest, and relationships between participants (strangers, colleagues, partners) (see Rajaram and Pereira-Pasarin (2010) for a review). Yet what about factors like discussion style, competitiveness, hierarchy, and group norms? It would be valuable to capture experiences closer to reality. Relatedly, it may be worth relaxing the degree to which the experimental conditions are controlled, and observing real-world collaboration to search for new insights: what happens with collaborative recall versus nominal recall in actual brainstorming sessions of different types? Before running new studies, it is also worth integrating current results into a comprehensive computational model: our results from Chapter 1, for example, do not fit into the current collaborative memory framework, and other patterns also likely need to be integrated.

There are similar limitations and directions for future work for the other problem areas. In Chapter 2, we asked "How do we decide what is best when people have different preferences?" There is an interesting distinction between what people say should be done and what people actually do (Bostyn et al., 2018; FeldmanHall, Dalgleish, et al., 2012; FeldmanHall, Mobbs, et al., 2012; Tassy et al., 2013). We analyzed what choices people selected for an uninvested third-party to make, which was a more nuanced and likely more accurate description of what people think should be done compared to self-reports (or placing participants in positions of self-interest). However, it would be interesting to probe what people say should be done in our experiment and compare those results with our findings in Chapter 2, especially since self-report is a common probe for "what people think should be done." Turning to a second future direction, we limited the scope of the study in Chapter 2 in various ways, such as not analyzing negative utilities. An immediate future direction is to parametrically extend our study, using the existing paradigm. More broadly, answering the question of what people think is best when people have different preferences will require analyzing many more scenarios. Mapping the distribution of normative preferences for other-benefitting decision-making is likely to be time-consuming, assuming the results follow some of the trends that appear in self-interested distribution, with differences observed across cultures (Rochat et al., 2009), situation, and context (see Güroğlu et al. (2010) on the importance of intent and alternative options on fairness in the ultimatum game). Yet people have an intuitive understanding of "what is correct," both on the individual level and also as a shared sense across humans (e.g. norms that an uninvested third party should try to reduce deaths, "commonsense morality" from Finkel et al. (2001)). As Chapter 2 illustrates, mathematical models can capture both the individual variation in people's beliefs, and the shared norms. It seems worthwhile to continue work in this direction: seeking to understand what people think is best in a diverse population, using computational models to ground the hypotheses and capture people's revealed preferences. Excitingly, the artificial intelligence community seems enthused by this line of thinking. Both Srivastava et al. (2019) and Yaghini et al. (2019) use mathematical models to capture contextual human perceptions of fairness, portending future exploration of intuitive human understandings of what is good.

In Chapter 3, we addressed a narrow version of the question "How do we make inferences about others' preferences from their actions?", and many future directions surface from the narrowness of our inquiry. In our case "action" meant "response time before making a decision," but one can imagine that "action" can be grounded out in concepts as broad as "any language utterance" or "any decision." An array of research questions opens from there. On the other hand, we can also ask what questions are sparked from our specific study. Our work is interesting in that it requires theory of mind: we asked what information people could infer about someone given they'd seen them pause while making a decision, requiring thinking about other people's thinking. It is fascinating that researchers can build computational models that conduct this type of reasoning, and that these inverted generative models can capture cognitive computations that often occur without our explicit awareness (e.g. Ong et al., 2019). Work that extends our paradigm could build in additional levels of inference (if Alice takes a while to make a decision, Bob infers that Alice has similar preferences for the two items, and Carol infers Bob infers this but knows that Alice is actually just flustered because of Bob watching her...). Future work could also continue building theory of mind models in domains outside of response time: teaching (Shafto et al., 2014), evaluating deception (Shafto et al., 2012; Sobel & Kushnir, 2013), pragmatics (Goodman & Frank, 2016), and many other areas (Jara-Ettinger, 2019; Shum et al., 2019) benefit from inverse generative models describing inherent cognitive algorithms.

As a final future direction, computational cognitive models of social collaboration have the potential to improve artificial intelligence systems aimed at promoting human welfare. For the work in Chapter 1 to be applicable to promoting human welfare, we will first need a stronger model of the mechanisms of collaborative memory. Following this, people could be nudged in the correspondingly beneficial direction, potentially by using assistive artificial intelligence systems to support environments that would improve human recall. Depending on the research results, those interventions could include adding natural language processing bots to contribute to a discussion, or gently nudging people away from first discussing as a group. Chapter 2's work is useful towards building artificial intelligence systems that act in accordance with human preferences, and informing human decision-makers about shared human preferences. However, for this work to be useful to artificial intelligence systems, much more empirical work is needed to understand the underlying algorithms driving human moral intuitions, and thus contribute to artificial intelligence systems reasoning about humans appropriately. Chapter 3 describes a model of how people make inferences about others' preferences from response times. Many people implicitly have this knowledge, but others (e.g. those with autism) may benefit from learning about it explicitly. Moreover, the greater application in my mind is the contribution to theory of mind algorithms. Theory of mind, in the sense of being able to infer others' mental states (goals, beliefs, etc.) from others' actions, may be valuable in building good human-artificial intelligence collaborations (Fisac et al., 2020), since social inference is so useful in seamlessly understanding, interacting with, and assisting people. (See Dafoe et al. (2020) for a "Cooperative AI" framework, in which theory of mind models are situated within the cooperative capability of "understanding of [other agents, their beliefs, incentives, and capabilities].") However, when discussing ideas like preference inference, we should also ensure that the work is applied in alignment with human values. We would probably want our artificial intelligence systems to infer our preferences if the system is unsure what to do in a moral situation or if we need help; we probably do not want our artificial intelligence systems to become so adept at personal marketing that we spend money we do not have. To summarize, formalizing and testing models of social collaboration is a fruitful approach for developing concrete and concise representations of the mind's computations, and also offers intriguing opportunities to improve collaboration between people, and between people and assistive artificial intelligence. Throughout the development of these artificial intelligence applications, we should ensure that they are serving human goals.

Overall, I have described work from three problem areas relevant to understanding and building better models of social collaboration. These problem areas are diverse, illustrating the potential of the space for future study and serving as exemplars for how formulating and testing computational cognitive models can advance our understanding of how people collaborate. One of the great complexities of social collaboration is how simple it feels when we do it, compared to how hard we find it to describe. My hope is that this work contributes to our described understanding of collaboration, a fascinating venture in and of itself, and also will contribute to better interactions between humans and human and artificial intelligence systems alike.

References

Baker, C. L., Jara-Ettinger, J., Saxe, R., & Tenenbaum, J. B. (2017). Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, *1*(4), 1–10. https://doi.org/10.1038/s41562-017-0064

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition*, *113*(3), 329–349. https://doi.org/https://doi.org/10.1016/j.cognition.2009.07.005

Bostyn, D. H., Sevenhant, S., & Roets, A. (2018). Of mice, men, and trolleys: Hypothetical judgment versus real-life behavior in trolley-style moral dilemmas. *Psychological science*, *29*(7), 1084–1093.

Chandrasekaran, B., & Conrad, J. M. (2015). Human-robot collaboration: A survey. *SoutheastCon 2015*, 1–8.

Chater, N., Tenenbaum, J. B., & Yuille, A. (2006). Probabilistic models of cognition: Conceptual foundations. *Trends in cognitive sciences*, *10*(7), 287–291.

Colman, A. M. (1995). *Game theory and its applications in the social and biological sciences*. Psychology Press.

Colman, A. M. (2003). Cooperation, psychological game theory, and limitations of rationality in social interaction.

Dafoe, A., Hughes, E., Bachrach, Y., Collins, T., McKee, K. R., Leibo, J. Z., Larson, K., & Graepel, T. (2020). Open Problems in Cooperative AI. *arXiv preprint arXiv:2012.08630*.

Engelmann, D., & Strobel, M. (2004). Inequality aversion, efficiency, and maximin preferences in simple distribution experiments. *American Economic Review*, *94*(4), 857–869. https://doi.org/https://doi.org/10.1257/0002828042002741

Fehr, E., & Schmidt, K. M. (2006). The economics of fairness, reciprocity and altruism–experimental evidence and new theories. *Handbook of the Economics of Giving, Altruism and Reciprocity*, *1*, 615–691.

FeldmanHall, O., Dalgleish, T., Thompson, R., Evans, D., Schweizer, S., & Mobbs, D. (2012). Differential neural circuitry and self-interest in real vs hypothetical moral decisions. *Social cognitive and affective neuroscience*, *7*(7), 743–751.

FeldmanHall, O., Mobbs, D., Evans, D., Hiscox, L., Navrady, L., & Dalgleish, T. (2012). What we say and what we do: The relationship between real and hypothetical moral choices. *Cognition*, *123*(3), 434–441.

Finkel, N. J., Harre, R., & Lopez, J.-L. R. (2001). Commonsense morality across cultures: Notions of fairness, justice, honor and equity. *Discourse Studies*, *3*(1), 5–27.

Fisac, J. F., Gates, M. A., Hamrick, J. B., Liu, C., Hadfield-Menell, D., Palaniappan, M., Malik, D., Sastry, S. S., Griffiths, T. L., & Dragan, A. D. (2020). Pragmatic-pedagogic value alignment. *Robotics Research* (pp. 49–57). Springer.

Fong, T., Thorpe, C., & Baur, C. (2003). Collaboration, dialogue, human-robot interaction. *Robotics Research* (pp. 255–266). Springer.

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*(6084), 998–998.

Gates, M. A., Suchow, J. W., & Griffiths, T. L. (2017). Empirical tests of large-scale collaborative recall. *Proceedings of the 39th Annual Conference of the Cognitive Science Society*.

Gates, V., Griffiths, T. L., & Dragan, A. D. (2020). How to Be Helpful to Multiple People at Once. *Cognitive Science*, *44*(6), e12841.

Gershman, S. J., Horvitz, E. J., & Tenenbaum, J. B. (2015). Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science*, *349*(6245), 273–278.

Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in cognitive sciences*, *20*(11), 818–829.

Griffiths, T. L. (2015). Manifesto for a new (computational) cognitive revolution. *Cognition*, *135*, 21–23.

Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., & Tenenbaum, J. B. (2010). Probabilistic models of cognition: Exploring representations and inductive biases. *Trends in Cognitive Sciences*, *14*(8), 357–364.

Griffiths, T. L., Kemp, C., & Tenenbaum, J. B. (2008). Bayesian models of cognition. In R. Sun (Ed.), *The Cambridge Handbook of Computational Cognitive Modeling*. Cambridge University Press.

Güroğlu, B., van den Bos, W., Rombouts, S. A., & Crone, E. A. (2010). Unfair? It depends: neural correlates of fairness in social context. *Social Cognitive and Affective Neuroscience*, *5*(4), 414–423.

Herreiner, D., & Puppe, C. (2007). Distributing indivisible goods fairly: Evidence from a questionnaire study. *Analyse & Kritik*, *29*(2), 235–258.

Jara-Ettinger, J. (2019). Theory of mind as inverse reinforcement learning. *Current Opinion in Behavioral Sciences*, *29*, 105–110. https://doi.org/10.1016/j.cobeha.2019.04.010

Jara-Ettinger, J., Gweon, H., Schulz, L. E., & Tenenbaum, J. B. (2016). The naïve utility calculus: Computational principles underlying commonsense psychology. *Trends in Cognitive Sciences*, *20*(8), 589–604. https://doi.org/10.1016/j.tics.2016.05.011

Jern, A., Lucas, C. G., & Kemp, C. (2017). People learn other people's preferences through inverse decision-making. *Cognition*, *168*, 46–64. https://doi.org/10.1016/j.cognition.2017.06.017

Kiley Hamlin, J., Ullman, T., Tenenbaum, J., Goodman, N., & Baker, C. (2013). The mentalistic basis of core social cognition: Experiments in preverbal infants and a computational model. *Developmental science*, *16*(2), 209–226.

Konovalov, A., & Krajbich, I. (2020). Decision times reveal private information in strategic settings: Evidence from bargaining experiments. *Available at SSRN 3023640*. https://doi.org/10.2139/ssrn.3023640

Lucas, C. G., Griffiths, T. L., Xu, F., Fawcett, C., Gopnik, A., Kushnir, T., Markson, L., & Hu, J. (2014). The child as econometrician: A rational model of preference understanding in children. *PloS One*, *9*(3). https://doi.org/10.1371/journal.pone.0092160

Luhmann, C., & Rajaram, S. (2015). Memory transmission in small groups and large networks: an agent-based model. *Psychological Science*, *26*, 1909–1917.

Ng, A. Y., & Russell, S. J. (2000). Algorithms for inverse reinforcement learning. In P. Langley (Ed.), *Proceedings of the 17th International Conference on Machine Learning* (pp. 663–670).

Ong, D. C., Zaki, J., & Goodman, N. D. (2019). Computational models of emotion inference in theory of mind: A review and roadmap. *Topics in cognitive science*, *11*(2), 338–357.

Perfors, A., Tenenbaum, J. B., Griffiths, T. L., & Xu, F. (2011). A tutorial introduction to Bayesian models of cognitive development. *Cognition*, *120*(3), 302–321.

Rajaram, S., & Pereira-Pasarin, L. (2010). Collaborative memory: Cognitive research and theory. *Perspectives on psychological science*, *5*(6), 649–663.

Rochat, P., Dias, M. D., Liping, G., Broesch, T., Passos-Ferreira, C., Winning, A., & Berg, B. (2009). Fairness in distributive justice by 3-and 5-year-olds across seven cultures. *Journal of Cross-Cultural Psychology*, *40*(3), 416–442.

Severinson-Eklundh, K., Green, A., & Hüttenrauch, H. (2003). Social and collaborative aspects of interaction with a service robot. *Robotics and Autonomous systems*, *42*(3-4), 223–234.

Shafto, P., Eaves, B., Navarro, D. J., & Perfors, A. (2012). Epistemic trust: Modeling children's reasoning about others' knowledge and intent. *Developmental science*, *15*(3), 436–447.

Shafto, P., Goodman, N. D., & Griffiths, T. L. (2014). A rational account of pedagogical reasoning: Teaching by, and learning from, examples. *Cognitive psychology*, *71*, 55–89.

Shum, M., Kleiman-Weiner, M., Littman, M. L., & Tenenbaum, J. B. (2019). Theory of minds: Understanding behavior in groups through inverse planning. *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*(01), 6163–6170.

Sobel, D. M., & Kushnir, T. (2013). Knowledge matters: How children evaluate the reliability of testimony as a process of rational inference. *Psychological Review*, *120*(4), 779.

Srivastava, M., Heidari, H., & Krause, A. (2019). Mathematical notions vs. human perception of fairness: A descriptive approach to fairness for machine learning. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2459–2468.

Tassy, S., Oullier, O., Mancini, J., & Wicker, B. (2013). Discrepancies between judgment and choice of action in moral dilemmas. *Frontiers in psychology*, *4*, 250.

Tenenbaum, J. B., Griffiths, T. L., & Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends in cognitive sciences*, *10*(7), 309–318.

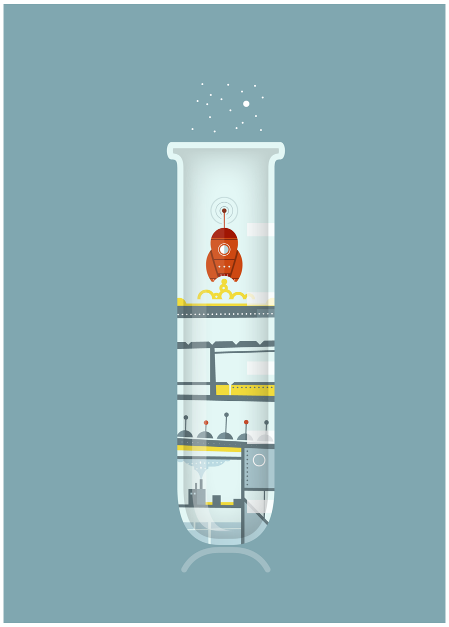
Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *science*, *331*(6022), 1279–1285.

Wang, D., Churchill, E., Maes, P., Fan, X., Shneiderman, B., Shi, Y., & Wang, Q. (2020). From Human-Human Collaboration to Human-AI Collaboration: Designing AI Systems That Can Work Together with People. *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–6.

Wilson, H. J., & Daugherty, P. R. (2018). Collaborative intelligence: humans and AI are joining forces. *Harvard Business Review*, *96*(4), 114–123.

Yaari, M. E., & Bar-Hillel, M. (1984). On dividing justly. *Social choice and welfare*, *1*(1), 1–24. https://doi.org/https://doi.org/10.1007/BF00297056

Yaghini, M., Heidari, H., & Krause, A. (2019). A Human-in-the-loop Framework to Construct Context-dependent Mathematical Formulations of Fairness. *CoRR*, *abs/1911.03020*. http://arxiv.org/abs/1911.03020



This thesis was typeset using LaTeX, originally developed by Leslie Lamport and based on Donald Knuth's TeX. The body text is set in 12 point Egenolff-Berner Garamond, a revival of Claude Garamont's humanist typeface. The above illustration, *Science Experiment 02*, was created by Ben Schlitter and released under CC BY-NC-ND 3.0. A template that can be used to format a PhD dissertation with this look & feel has been released under the permissive AGPL license, and can be found online at github.com/suchow/Dissertate or from its lead author, Jordan Suchow, at suchow@post.harvard.edu.