Title

Formalizing and Testing Computational Cognitive Models of Social Collaboration

Author

Gates, Vael

Publication Date

2021

Peer reviewed|Thesis/dissertation

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, Matt Elliot, Kealey McKown, Cindi Baker-Smith, Matt Elliot, 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, Jabob 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; T. Fong, Thorpe, & Baur, 2003; Severinson-Eklundh, Green, & Hüttenrauch, 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, Tenenbaum, & Yuille, 2006; Gershman, Horvitz, & Tenenbaum, 2015; Griffiths, 2015; Griffiths, Chater, Kemp, Perfors, & Tenenbaum, 2010; Griffiths, Kemp, & Tenenbaum,

2008; Perfors, Tenenbaum, Griffiths, & Xu, 2011; Tenenbaum, Griffiths, & Kemp, 2006; Tenenbaum, Kemp, Griffiths, & Goodman, 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, Jara-Ettinger, Saxe, & Tenenbaum, 2017; Baker, Saxe, & Tenenbaum, 2009; Jara-Ettinger, Gweon, Schulz, & Tenenbaum, 2016; Jern, Lucas, & Kemp, 2017; Kiley Hamlin, Ullman, Tenenbaum, Goodman, & Baker, 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; Ernst 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, Suchow, & Griffiths, 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 Introduction

Our memories depend on the people around us and are reshaped and reformed as we interact. Long-term couples, for example, learn how to effectively access each other's memories, remembering details they would not have recalled alone (Wegner, Erber, & Raymond, 1991), and groups collectively form memories that define their values (e.g., Hirst & Echterhoff, 2012). However, sometimes our memories are reshaped in such a way that we remember less together than we would apart, an effect that psychologists have been actively investigating.

Imagine a group of people trying to recall a list of words collaboratively. The group as a whole would recall some number of words. Next suppose that each individual had instead tried to recall the list of words alone. Though the group would likely have recalled a greater number of words than any individual, the group may have performed even better had they recalled the words individually and then mechanically combined the list, removing duplicates. Indeed, this precise comparison is often made in analyses of the well-established "collaborative memory" task, in which participants listen to a long list of items (often words) and then recall as many items as possible, either as a group or individually. The number of words recalled by the group is compared to the number of words recalled by the "nominal group": the summed list of an equivalent number of individuals, with duplicate words removed. In the collaborative memory paradigm, nominal groups routinely outperform collaborative groups, a finding called *collaborative inhibition*. This effect has been repli-

cated across many studies and variations on the paradigm (see Marion & Thorley, 2016; Rajaram & Pereira-Pasarin, 2010, for reviews).

The leading theory of collaborative inhibition is the retrieval disruption hypothesis (e.g., Basden, Basden, Bryner, & Thomas, 1997; Rajaram & Pereira-Pasarin, 2010). This hypothesis states that when initially listening to a wordlist, people form idiosyncratic representations of the words, influenced by factors such as the order of the words presented and their semantic relations. When recalling words alone, participants use their idiosyncratic representations to effectively recall the words. However, when placed in groups, other participants, who organized the words differently, disrupt a participant's recall, leading to reduced performance. This hypothesis predicts that when participants are encouraged to organize information in similar ways, collaborative inhibition will disappear. In fact, when participants are experts in their domain (Meade, Nokes, & Morrow, 2009) or are exposed to similarly ordered information (Finlay, Hitch, & Meudell, 2000), inhibition does not occur.

Psychologists have historically studied collaborative memory in small-scale experiments of dyads or triads in the lab. Bringing groups of participants into the lab to simultaneously recall information—often lists of words—is logistically complex, slow, and expensive. However, in today's world of online communities and internet forums, information is being recalled and exchanged with much larger groups than pairs or triads, and many of the findings from this small-scale work may not be applicable to our daily lives. As people form ever-larger social networks, understanding how information is recalled among large groups of people becomes increasingly important.

To address this lack of understanding of large groups, Luhmann and Rajaram (2015) developed an agent-based model to predict what performance might be like at larger group sizes, based on known factors from empirical small-scale experiments. Computational "agents" in these models follow algorithms for memory storage and retrieval, and can participate in simulated experiments interacting with many other agents.

Previously, it was a logistical impossibility to experimentally study collaborative memory at this kind of scale. However, using web-based crowdsourcing tools such as Amazon Mechanical Turk it is possible to recruit and organize hundreds of online participants into real-time interactive chatrooms. Here we use this new approach to translate the agent-based model of Luhmann and Rajaram (2015) into an empirical experiment, testing how people will recall and share information at scale.

1.2 Experiment 1: Collaborative inhibition in small and large groups

To examine the predictions of the retrieval disruption hypothesis for collaborative inhibition as group size increases, Luhmann and Rajaram (2015) developed an agent-based model that they used

to simulate a collaborative recall task. Agents encounter items one by one, increasing the activation of items at a rate determined by a parameter α while decreasing the activation of unrelated items at a rate determined by a parameter β, and then in a recall phase randomly take turns in which they have a probability γ of generating an item, with higher activated items being more likely to be retrieved and retrieved items decreasing the activation of associated items for all agents (details of the model appear in the Supplemental Materials).

This agent-based model predicts that collaborative inhibition will increase as group size increases from 1 to 4 participants, as has been suggested by the experimental literature (Basden, Basden, & Henry, 2000; Thorley & Dewhurst, 2007). The model then predicts that collaborative inhibition will continue to increase with group size, until nominal recall reaches ceiling performance as the disruption of idiosyncratic recall strategies is compensated for by sheer group size. From this point, collaborative inhibition will begin to decrease, since collaborative inhibition is the difference between nominal and collaborative recall.

When we simulate the model with parameter settings corresponding to our experimental task, it predicts nominal ceiling performance around a group size of 16.[1] Our experiment thus used group sizes of 2 through 16, meaning that the model predicts increasing collaborative inhibition up to the largest groups (Figure 1a). More precisely, it predicts an interaction between group size and recall method (collaborative or nominal), with the effect of recall method increasing with group size.

1.2.1 Methods

All experiments were IRB-approved by University of California, Berkeley, Committee for Protection of Human Subjects / Office for Protection of Human Subjects, Protocol ID: 2015-12-8227, Protocol Title: Culture-on-a-chip Computing: Testing Evolutionary Hypotheses through Large-Scale Behavioral Simulations.

[1] We present model predictions using identical parameter settings to Luhmann and Rajaram (2015) with $\alpha = 0.2$, $\beta = 0.05$ and $\gamma = 0.75$ except we use 60 recalled words rather than 40, because pilot studies indicated that participants were nearing ceiling performance when recalling fewer words (an analysis of the results of this change appears in the Supplemental Materials). Note also that when recalling 60 words, the model prediction of increased collaborative inhibition with group size appears robustly across parameter settings (see section Analyzing the Model below).

Participants

A total of 1,138 participants were recruited through Amazon Mechanical Turk. Participants would occasionally repeat the task, as they could choose to complete the task again on Amazon Mechanical Turk despite written advisement against this. A total of 134 participants repeated the experiment more than once (14.6% of participants), and 30.3% of the data was generated by these participants.[2]

Participants were excluded from the experiment if they did not complete a pre-experiment arithmetic task and they did not contribute words in the main experiment. There were two recall conditions in the experiment: collaborative and nominal. Sixteen participants were removed from the collaborative experiments for a total of 561 participants. Nominal groups were matched; thus 561 participants participated in the nominal experiments. The average ($\pm$ SD) number of participants in collaborative and nominal experiments was 15.2 ± 0.7 for groups of size 16, 7.6 ± 0.7 for groups of size 8, 4.0 ± 0.0 for groups of size 4, 3.0 ± 0.0 for groups of size 3, and 2.0 ± 0.0 for groups of size 2. (In group sizes of 4, 3, and 2, groups were only included if they contained the maximum number of participants. In the nominal condition, participants completed the experiments individually and then were only later matched to the appropriate group sizes, so all participants were included.) Participants were paid $1.00 for completing the task, plus bonuses for time spent waiting for other participants to arrive in a chatroom to begin the experiment ($5.00/hour).

Procedure

Participants were presented with a wordlist: 60 unrelated words each selected from a different category from Overschelde, Rawson, and Dunlosky (2004). (Example words include "diamond," "hour," and "uncle.") Each word was presented for two seconds. After seeing the list, participants completed an arithmetic filler task for 30 seconds before advancing to the recall task. Participants were placed in chatrooms alone or with other participants, and were encouraged to type as many of the words they had seen as possible. Participants were free to submit words at any time. Participants

[2] Participants participated an average of 1.22 times. The mean proportion of repeaters across group sizes (both nominal and collaborative) was as follows: group size of 2: 0.33 ± 0.35 (SD), group size of 3: 0.39 ± 0.31, group size of 4: 0.27 ± 0.21, group size of 8: 0.27 ± 0.20, and group size of 16: 0.27 ± 0.09. The participants who repeated the nominal experiments did not show improvement over time, despite having seen the same wordlists: the correlation between number of repetitions and number of words recalled was r=-0.03 (91 data points). Participants did improve across repetitions in the collaborative experiments (r=0.47, 89 data points). However, the proportion of repeaters within experiments was anti-correlated with group size: the correlation between experiments of group sizes 2, 3, 4, 8, and 16 and the proportion of repeaters was r=-0.086.

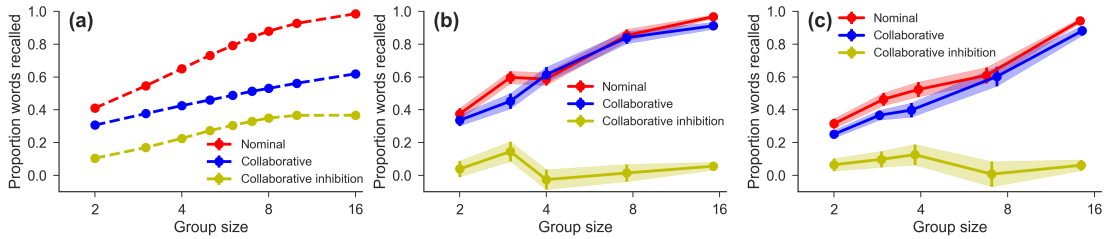


Figure 1.1: Effect of group size on collaborative inhibition. (a) Model predictions, produced using code from Luhmann and Rajaram (2015). (b) Experiment 1 behavioral results. The mean proportion of words (± SE) recalled in the nominal (red/grey) and collaborative (blue/black) conditions are shown; collaborative inhibition (yellow/light-grey) is the subtraction of these two values. Note that the horizontal axis is logarithmic. (c) Experiment 2 behavioral results.

were not told how many other participants were in the chatroom: their responses appeared in blue font, and responses from all others appeared in black. They saw all previous words entered and were not permitted to submit any word that had already been submitted (by anyone in the group). This choice—that any words already present on the group list were not redisplayed—was made to encourage participants to read others' submitted words, and because it more closely matched the lab-based version of the collaborative memory paradigm, in which verbal recall creates social pressure to not repeat words. There was no time limit for the recall task.

Groups contained 2, 3, 4, 8, or 16 participants. For each recall method (nominal or collaborative), 48 groups of 2 were analyzed; 32 groups of 3 were analyzed; 24 groups of 4 were analyzed, and 12 groups of 8 and 16 were analyzed. The sample sizes for group sizes 2, 3, and 4 were consistent with and generally larger than those in previous experiments (Marion & Thorley, 2016). An unbalanced design was used to ensure a similar number of participants for each group size. This was a 2×5 design, crossing recall method by group size. The dependent variable was the number of words recalled for each group. In the nominal recall condition, participants recalled words alone (and were not informed their recall lists would later be pooled with other participants'). To analyze the results after the experiment, nominal participants' recall lists (with redundant words removed) were added together according to the appropriate group size. In the collaborative recall condition, groups of participants were placed in chatrooms and recalled together. Recalled words that had not been on the original lists were marked as incorrect and not included.

Participants completed a post-experiment questionnaire, in which we asked about participants' engagement in the task (1-10), the perceived difficulty of the task (1-10), and had a textbox for any other comments.

	t	p	d	M	*95% CI*
Group size 2	$t(94)$=0.78	0.44	0.16	2.33	[-3.53, 8.19]
Group size 3	$t(62)$=2.34	0.022	0.60	8.69	[1.40, 15.98]
Group size 4	$t(46)$=-0.42	0.67	0.13	-1.58	[-8.94, 5.77]
Group size 8	$t(22)$=0.25	0.80	0.11	0.83	[-5.74, 7.41]
Group size 16	$t(22)$=2.17	0.041	0.93	3.33	[0.29, 6.38]

Table 1.1: Two-sample t-test results for each group size in Experiment 1. Alongside t-test and effect size (Cohen's d) statistics we report the mean difference (M) in words recalled between the nominal and collaborative groups and associated confidence intervals (CI).

1.2.2 Results

The collaborative inhibition effect is most reliably observed in triads (Rajaram & Pereira-Pasarin, 2010), and we replicated this effect in our behavioral data at group size 3: $t(62)$=2.34, p=0.022, effect size d=0.60, independent two-sample t-test (Figure 1b, Table 1.1).[3] Collaborative inhibition in the literature is frequently but not always observed in pairs (Rajaram & Pereira-Pasarin, 2010), but we did not observe this effect in this group size (p=0.44, d=0.16). There are only two studies known to the authors examining tetrads (Basden et al., 2000; Thorley & Dewhurst, 2007). Collaborative inhibition effects were observed in these studies, but we did not observe this effect at group size of 4 (p=0.67, d=0.13).

Given previous observations of the collaborative inhibition effect for small group sizes, Luhmann and Rajaram (2015) extrapolated and hypothesized that the effect of collaborative inhibition would be expected to increase for larger group sizes until nominal group recall reached 100% (Figure 1a). There was no significant collaborative inhibition effect at group size of 8 (p=0.80, d=0.11). However, we did find a collaborative inhibition effect at group size 16 (p=0.041, d=0.93, note that effect size is especially large due to small variance in word recall because of ceiling effects).

Overall, using a between-participants two-way unbalanced ANOVA, we did not observe a statistically significant main effect of recall method, $F(1,246)$=2.03, p=0.16, η^2=0.0045: nominal and collaborative groups did not recall significantly different numbers of words when results from all groups were combined (Figure 1b). We did observe an expected main effect of group size, $F(4,246)$=49.08, p<0.0001, η^2=0.44, in that larger group sizes increased word recall. Critically, there was no interaction effect between recall method and group size, $F(4,246)$=1.15, p=0.33, η^2=0.010. This was the

[3] $\alpha = 0.05$ for all planned comparisons to follow. Unlike previous studies, we investigate the collaborative inhibition effect at multiple group sizes, while previous studies used $\alpha = 0.05$ at a single group size, justifying its use here.

key prediction of the model proposed by Luhmann and Rajaram (2015).

We conducted a power analysis to determine whether our sample sizes were sufficient to detect effects as large as those predicted by the model. We generated 1000 simulations of the entire experiment using the parameter settings from Luhmann and Rajaram (2015) with 60 words. In each simulation, the model was run a number of times matching the sample sizes used in the experiment,[4] and we conducted an ANOVA testing for a main effect of recall, a main effect of group size, and an interaction effect. In 1000/1000 simulations, the model had p<1e-40 for all these effects. We conclude that if our participants had behaved like the model predicted, we would have had enough power to detect these effects.

1.3 Experiment 2: Preregistered replication with a modified task

Because the main finding from Experiment 1 was a null effect for the predicted interaction we conducted a preregistered replication (Experiment 2). In the replication, we modified the task structure to remove factors that may have suppressed memory interference. This experiment was preregistered (planned methods and analyses) with the Open Science Framework (https://osf.io/7ht3g).

1.3.1 Methods

Participants

A total of 1,135 participants were recruited through Amazon Mechanical Turk. Participants were not permitted to repeat the task. As in Experiment 1, participants were excluded from the experiment if they did not complete the pre-experiment arithmetic task and they did not contribute words in the main experiment. 575 participants were recruited and 36 participants were excluded from the collaborative conditions, for a total of 539 participants. The average ($\pm$ SD) number of participants in the collaborative conditions was 2.0 $\pm$ 0.0 for groups of size 2, 2.9 $\pm$ 0.3 for groups of size 3, 3.7 $\pm$ 0.5 for groups of size 4, 7.3 $\pm$ 0.6 for groups of size 8, and 14.5 $\pm$ 1.4 for groups of size 16. 560 participants were recruited and 23 participants were excluded from the nominal conditions, for a total of 537 participants. The average ($\pm$ SD) number of participants in the nominal conditions was 2.0 $\pm$ 0.0 for groups of size 2, 3.0 $\pm$ 0.2 for groups of size 3, 3.9 $\pm$ 0.3 for groups of size 4, 6.8 $\pm$ 0.8 for groups of size 8, and 14.3 $\pm$ 1.2 for groups of size 16. (Across both the collaborative and nominal conditions, groups with less than two participants were re-run.) Participants were paid

[4] Sample sizes: 48 groups for group size 2, 36 groups for group size 3, 24 groups for group size 4, and 12 groups for group sizes 8 and 16.

$1.00 for completing the task, plus bonuses for time spent waiting for other participants to arrive in a chatroom to begin the experiment ($5.00/hour) and bonuses for the number of correct words they recalled and participants who transmitted words to them recalled ($.015/word).

Procedure

Participants observed 60 words, randomly selected for each group of participants across all conditions, drawn from a list of 200 words from Overschelde et al. (2004). (Example words include "adverb,""pine," and "dollar.") Each group had a different wordlist presented, and within each group, the order of presented words in the wordlist was selected randomly for each participant. Each word was presented for two seconds. After seeing the list, participants completed a 30-second-long arithmetic filler task before advancing to the recall task. Participants were placed in chatrooms alone (the nominal recall condition) or with other participants (the collaborative recall condition), and were encouraged to type as many of the words they had seen as possible.

As in Experiment 1, participants were not permitted to submit any word that had previously been submitted, to increase similarity to the lab-based version of the collaborative memory paradigm in which verbal recall creates social pressure to not repeat words. Moreover, words that were submitted by others in the chatroom, and not by the participant, were read aloud to the participant, again to increase similarity to the lab-based task. Words were read aloud by a text-to-speech computer algorithm that automatically converted participant text during the experiment. To ensure that participants could understand the computerized speech and that their audio was working, before the task began participants had to correctly type in an isolated phrase read aloud by this computerized voice. In the nominal conditions, no words were read aloud, since no other participants were present in the chatroom. There was no time limit for the recall task.

To increase similarity to the specific methodology of Luhmann and Rajaram (2015), participants in the collaborative conditions were not permitted to submit words whenever they wished (as was true in the nominal condition). Rather, participants took turns to submit words and did so in random order, following Luhmann and Rajaram (2015). (Both turn-based and free-recall structures are common collaborative memory paradigms, as described in Marion and Thorley (2016), Rajaram and Pereira-Pasarin (2010).) Participants took part in "rounds," in which each participant had the option to submit one word during their five-second turn, and turn order was determined randomly within each round. Participants could either submit a word that had not already been submitted, press a "Pass" button, or wait for their 5-second turn to elapse on each turn. When participants exited the study, they were no longer included in the turn order for the remaining participants.

	t	*p*	*d*	*M*	*95% CI*
Group size 2	$t(94)$=-1.64	0.10	0.34	3.85	[-0.76, 8.47]
Group size 3	$t(62)$=-1.95	0.056	0.50	5.84	[-0.054, 11.74]
Group size 4	$t(46)$=-1.96	0.056	0.58	7.54	[-0.035, 15.12]
Group size 8	$t(22)$=-0.086	0.93	0.037	0.42	[-9.17, 10.00]
Group size 16	$t(22)$=-1.76	0.093	0.75	3.67	[-0.48, 7.81]

Table 1.2: Two-sample *t*-test results for each group size in Experiment 2. Alongside *t*-test and effect size (Cohen's *d*) statistics we report the mean difference (M) in words recalled between the nominal and collaborative groups and associated confidence intervals (CI).

As in Experiment 1, groups contained 2, 3, 4, 8, or 16 participants. For each recall method (nominal or collaborative), 48 groups of 2 were analyzed; 32 groups of 3 were analyzed; 24 groups of 4 were analyzed, and 12 groups of 8 and 16 were analyzed. The sample sizes analyzed were on par with and generally larger than those in other collaborative memory experiments (Marion & Thorley, 2016).

Participants completed a post-experiment questionnaire, in which we asked about participants' engagement in the task (1-10), the perceived difficulty of the task (1-10), and had a textbox for any other comments.

1.3.2 Results

Recall that in this experiment, we would observe a collaborative inhibition effect if the nominal groups recalled more words than the collaborative groups. For all group sizes in Experiment 2, we observed that the nominal groups did indeed recall more words than the collaborative groups in absolute number, but this effect was small (Figure 1c). When results from all groups were combined, nominal and collaborative groups recalled significantly different numbers of words: in a between-participants two-way unbalanced ANOVA, we observed a main effect of recall method, $F(1,246)$=6.47, p=0.012, η^2=0.013. We ran independent 2-sample *t*-tests for group sizes 2, 3, 4, 8, and 16, with α=0.05 for all planned comparisons shown (Table 1.2).

To summarize, we did not observe a statistically significant collaborative inhibition effect for group size 2 (p=0.10, effect size d=0.34). We saw marginally significant effects at group size 3 (p=0.056, d=0.50) and group size 4 (p=0.056, d=0.58) (Figure 1c). Consistent with previous work showing collaborative inhibition at small group sizes, the effect sizes for group sizes of 3 and 4 were similar to the adjusted estimate of overall mean effect size (d=0.56) for collaborative inhibition (calculated in a meta-analysis by Marion and Thorley (2016), using studies of group sizes 2-4 and compensating

for publication bias) and the effect size for group size 3 was similar to that in Experiment 1 (d=0.60). There were no statistically significant differences at group size 8 (p=0.93, d=0.037), nor at group size 16 (p=0.093, but note the large effect size of d=0.75 due to small variance in the proportion of words recalled because of ceiling effects). The number of words recalled did increase with group size, and in a between-participants two-way unbalanced ANOVA, we observed the expected main effect of group size, $F(4,246)$=56.38, p=1.0e-33, η^2=0.47. However, once again we found no interaction effect between recall method and group size, $F(4,246)$=0.47, p=0.76, η^2=0.0039, contrary to model predictions from Luhmann and Rajaram (2015).

1.4 Analyzing the model

The results of our experiments are not consistent with the predictions presented in Luhmann and Rajaram (2015). However, one could argue that if we had altered the parameters of the model, we would have seen patterns that were more compatible with our empirical results. Here we conduct two analyses, showing that across parameter settings, the empirical results do not match model predictions.

Our first analysis asked what predictions would be possible with the model, given the constraint of trying to match the empirical results as closely as possible. We thus did a grid search over the model's parameters, minimizing the squared error between the mean nominal and collaborative results for each group size.

Even for the best-fitting model, the model predictions do not qualitatively match the empirical results for Experiment 1 (Figure 1.2) or Experiment 2 (Figure 1.3). Specifically, the model is unable to predict a collaborative inhibition effect at small group sizes (as has been shown in the literature and our experiments) even with parameters optimized for reducing the discrepancy between the model and empirical results.

Our second analysis asked a different question: how often would we see model trends that fit the empirical data if we used a broad array of parameter settings? One way to operationalize the differences between the model predictions and empirical results is to check whether there is a smaller or larger collaborative inhibition effect for groups sizes 2, 3, and 4 compared to group sizes 8 and 16. The model predicts a smaller collaborative inhibition effect for group sizes 2, 3, and 4. In our empirical data, we observe a larger collaborative inhibition effect for group sizes 2, 3, and 4. This is one of the most qualitatively salient differences between the model and our results.

To examine the range of parameter values that might produce this effect, we conducted a grid search centered on the parameters in Luhmann and Rajaram (2015) (α=0.20, β=0.05, γ=0.75,

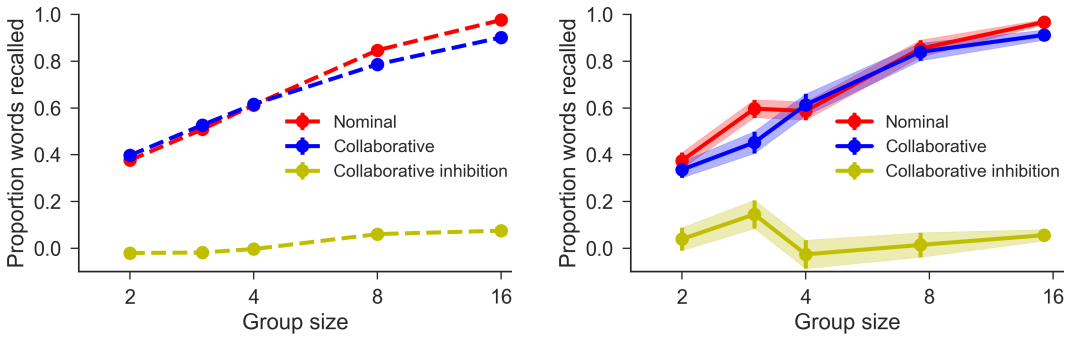


Figure 1.2: Best-fit model predictions and behavioral results for Experiment 1. (**Left**) Model predictions with the best-fitting parameters (α=0.10, β=0.07, γ=0.55, number of rounds=23), 1000 simulations. (Compare to the Luhmann and Rajaram (2015) model parameters of α=0.20, β=0.05, γ=0.75, number of rounds=20.) (**Right**) Behavioral results for Experiment 1.

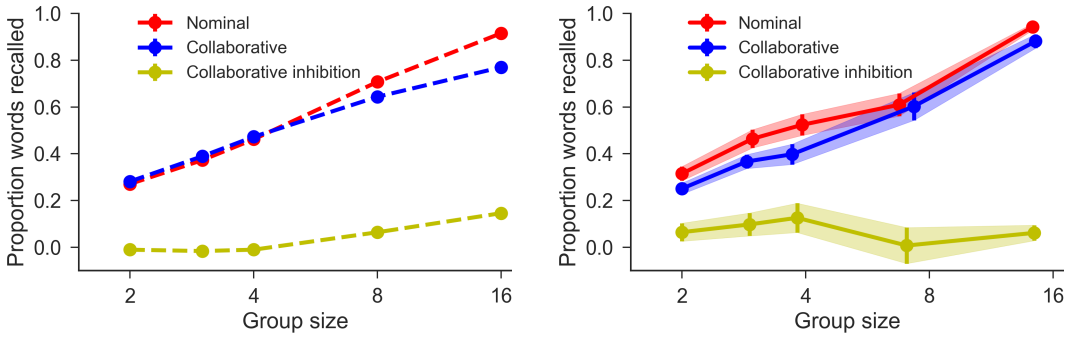


Figure 1.3: Best-fit model predictions and behavioral results for Experiment 2. (**Left**) Model predictions with the best-fitting parameters (α=0.12, β=0.03, γ=0.39, number of rounds=22), 1000 simulations. (Compare to the Luhmann and Rajaram (2015) model parameters of α=0.20, β=0.05, γ=0.75, number of rounds=20.) (**Right**) Behavioral results for Experiment 2.

number of rounds=20), with 60 words. We ran 1000 simulations for each of the following parameter combinations: α=[0.1,0.2,0.3]; β=[0,0.05,0.10]; γ=[0.55,0.65,0.75,0.85,0.95]; number of rounds=[10,20,30]. For each parameter combination, we tested whether the model would predict a larger collaborative inhibition effect for the average recall of group sizes 2, 3, and 4 compared to the average recall of group sizes 8 and 16, to match what we observed empirically.[5] We also placed the constraint of needing at least some collaborative inhibition for group sizes 2, 3, and 4: specifically, the average collaborative inhibition effect had to be greater than 0.01 (109/135 of the parameter combinations met this criteria). We placed this constraint because we considered it necessary for the model to predict collaborative inhibition at small group sizes to match our results.

In 9/109 of the parameter combinations, the model predicted a larger collaborative inhibition effect for group sizes 2, 3, and 4 than for group sizes 8 and 16, as was true in our empirical data. However, all of those nine parameter combinations included setting the number of rounds to 30 (Figure 1.4). As rounds increase, agents have more chances to recall words and reach ceiling performance more rapidly. Since the collaborative inhibition effect is the difference between nominal and collaborative group recall, if ceiling performance for both the nominal and collaborative groups is reached, collaborative inhibition decreases. These simulation results (Figure 1.4) show that if ceiling performance had not been reached as rapidly, collaborative inhibition would have increased rather than decreased with group size. Thus, even though for these parameter settings the model predicts decreased collaborative inhibition at larger group sizes, this effect can be explained by ceiling performance effects. This is not the pattern of results we see in our empirical data, in which larger group sizes show decreased collaborative inhibition without ceiling-level performance.

The model's predictions are even more inconsistent with our results when considering parameter combinations in which the number of rounds was 10 or 20: the model predicts that collaborative inhibition will increase at larger group sizes, rather than decrease. We thus observed that the model predictions did not qualitatively match empirical results across a variety of parameter settings, and when the model did predict larger collaborative inhibition at small group sizes, this was due to ceiling performance effects.

[5] In the empirical results, the collaborative inhibition effect as measured as the difference of proportion of words recalled was the following: the average of group sizes 2, 3, and 4 was 0.05 (Experiment 1) and 0.10 (Experiment 2), and the average of group sizes 8 and 16 was 0.03 (Experiment 1) and 0.03 (Experiment 2).

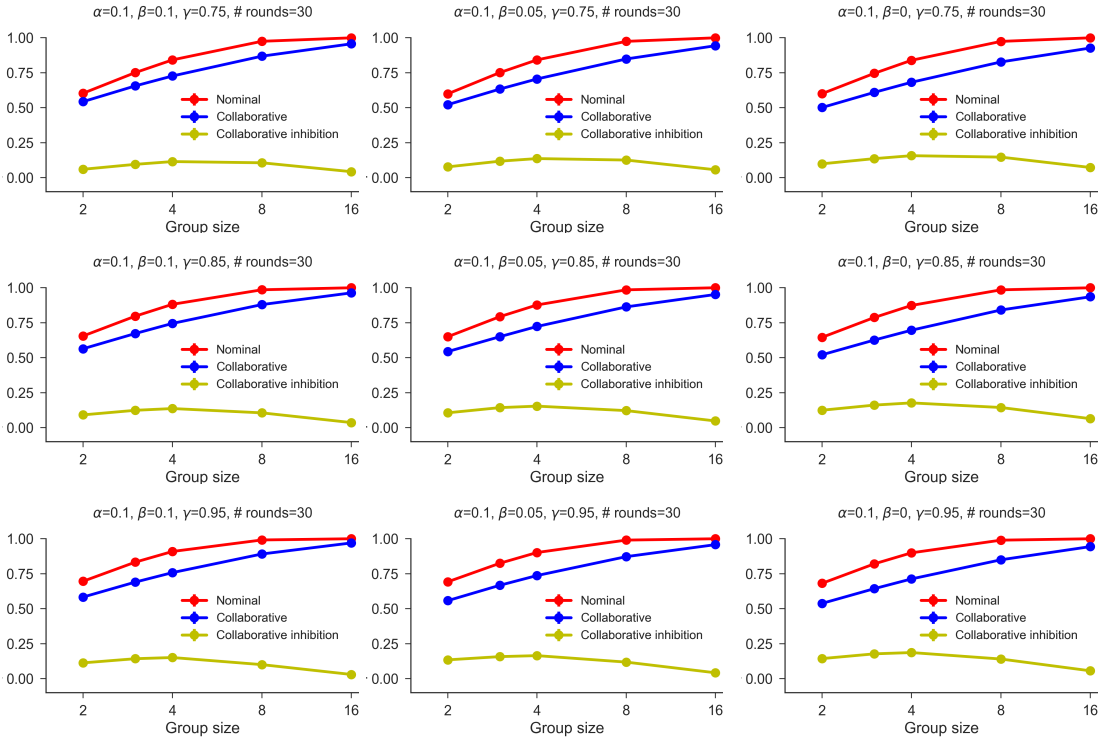


Figure 1.4: Parameter settings in which the model predicted a larger collaborative inhibition effect for the average recall of group sizes 2, 3, and 4 compared to the average recall of group sizes 8 and 16, constrained to simulations where there was an average collaborative inhibition effect for group sizes 2, 3, and 4 greater than 0.01. We ran 1000 simulations for each of the following parameter combinations: α=0.1, 0.2, 0.3; β=0, 0.05, 0.10; γ=0.55, 0.65, 0.75, 0.85, 0.95, number of rounds=10, 20, 30.

1.5 General Discussion

The critical prediction that Luhmann and Rajaram (2015) made is that collaborative inhibition—an effect found robustly at smaller group sizes—would be even greater with group sizes 8 and 16. Under their model, researchers would expect to see an interaction effect between recall method (collaborative or nominal) and group size, reflecting increased collaborative inhibition as group size increases. We did not observe this interaction effect in our data, and believe we had enough power to detect this effect had participants behaved as the model predicted. We now consider possible interpretations of these results.

One possibility is that our online task differed from laboratory studies in ways that had a meaningful impact on the results. First, we note that the online task produced results consistent with those from the laboratory – we observed collaborative inhibition at group size 3 in Experiment 1 and had marginally significant evidence for an effect at group sizes 3 and 4 (p=0.56) as well as a main effect of collaborative inhibition across groups in Experiment 2, and the size of these effects corresponds to those suggested by meta-analysis of laboratory experiments (Marion & Thorley, 2016). Second, we note that all of the factors relevant to the agent-based model – and to the broader theory of retrieval disruption – are present in our online tasks, which closely mimic the structure of the model. If there are critical features of the laboratory setting that are missing from our task, they are equally absent from the model. Finally, our online tasks – in which people interacted in chatrooms – are increasingly representative of the contexts in which people interact with large groups, making them a reasonable object of study in themselves.

Another possibility is that while retrieval disruption is the dominant force in smaller groups, additional factors come into play. For example, it might be the case that intense competition as group size increases drive greater recall. Alternatively, at larger group sizes, groups may need to adopt different representational strategies less prone to disruption. To determine what these different mechanisms would be, we will need controlled experiments aimed at developing additional cognitive, social, or motivational hypotheses explaining the unexpected lack of consistent collaborative inhibition at large group sizes.

Participant anecdotes can hint towards what makes large groups distinct. In post-Experiment 1 questionnaires, for group size 16, participants reported time pressure and competitive stress from many people racing to submit responses at once. In smaller groups, the number of responses was less overwhelming and more cooperative. In Experiment 2, participants took timed turns rather than experiencing free recall, but in larger groups people had to comparatively wait much longer for their turns compared to in smaller groups. These are examples of factors that are comparatively

minor in small group collaborative recall, but were qualitatively salient in large group collaborative recall.

In an increasingly interconnected world, it is essential to understand how information is shared and remembered by interacting groups of people. Methods like agent-based modeling provide a starting point for developing hypotheses about groups that are larger than we can normally study in the laboratory. However, crowdsourcing and computer-controlled recruitment offer new tools that we can use to evaluate these hypotheses and extend the scope of our empirical results. Our findings raise the possibility that there may be additional factors that influence collaborative recall in large group sizes, illustrating the importance of translating agent-based models based on lab studies into online studies that can improve our understanding of the interactive behavior of large groups of people.

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 op-

timal, 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, Griffiths, & Dragan, 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 Introduction

Consider a schoolteacher trying to plan a field trip. Some students learn best kinetically, others enjoy verbal challenges, and some would be overwhelmed by new locations. If there's no one action that will maximally satisfy everyone, what is the teacher to do? Now consider a concerned citizen determining how to donate their money. They furrow their brow, staring at the screen: donate to the most needy, the organization where their donations will be matched, or the recipient with the tightest time constraints? Next consider a high-level government worker, puzzling over who to prioritize, faced with an array of programs that will benefit everyone to some extent but some more than others. Finally consider an autonomous household robot, trying to determine what to make for the main family meal with one kid who only wants to eat orange foods and one parent who is vegan. What should all of these decision-makers do?

In today's society, as preference-aggregation problems become more complex, we can turn to tools from computer science to address them. If modern artificial intelligence systems know about the preferences of their recipients, they can use vast computational resources to optimize: airplane

scheduling and ride-share services are examples of using artificial intelligence to optimize many people's preferences. However, these consumer services optimize based on the principles of first-come-first-serve, more resources for more money, and efficiency. Individuals put in a bid and they receive some service. But in other types of situations, people advocate for others rather than themselves. People choose which organizations to donate money towards, and governments strive to serve their entire populations with aid programs. These situations are just as complex and deserving of computational analysis, but they will need a different optimization procedure: one that likely incorporates fairness.

To develop artificial intelligence tools to help solve coordination problems, we need to know what the ground truth is. What internal algorithms do people use to make decisions that benefit others? People don't always act in the ways they say they do, but by observing their ground truth behavior, we can gain a quantitative understanding of what behaviors people think are right. In computer science, there are existing tools for inferring the values, preferences, and utilities that motivate people's actions and choices. One standard method, inverse reinforcement learning (Ng & Russell, 2000), has been used to infer utility functions from behavior in both the robotics community (e.g. Kuderer, Gulati, and Burgard (2015), in which people's driving styles were inferred from demonstrations) and the computational cognitive science community (e.g. Baker et al. (2009), in which intended goals were inferred from partially completed paths). There are thus established methods to learn human preferences from actions: given examples of a person's driving, we are advancing in our ability to tell an autonomous vehicle how to bring them to the store; given examples of a person tidying up their home, we are becoming able to tell a robot what to clean. Yet these methods only apply to a single person's preferences at a time. What should be done when there is more than one person, and there are a combination of utilities to optimize? Should an artificial intelligence sum up all of its users' utilities, or would it be more "fair" to minimize the upset of the unhappiest member of the group? When we have assistive robots who must act as decision-makers themselves, how should they be programmed? We think the first step to answering these questions is to quantitatively determine how people—with their intuitions about fairness and efficiency and what is good—behave.

In this paper, we set out to map the human sense of compromise, and the challenge of the benign dictator: how to solve the age-old problem of acting in *everyone's* best interest without a vested interest of one's own. It's a challenging problem that has puzzled philosophers, arbitrators, and governors for centuries. The question of what should be done when people disagree is essential, and to build tools that can help solve it, we must understand the ground-truth of what we want done. This problem becomes increasingly important not only to advise decision-makers, but to develop artificial intelligence systems to help decision-makers and eventually create artificial intelligence agents

that can make choices that match what we do ourselves. In the face of this unsolved problem, we turn to psychology, developing a quantitative model of what people think should be done when an agent can take only one action that brings different utilities to different people.

Many researchers have thought about these questions. Their work often falls under the heading of *fair allocation*, which investigates how items should be divided among people. The study of fair allocation and fair division (Brams & Taylor, 1996; Konow, 2003) spans many fields, including those of social choice (Gaertner & Schokkaert, 2012; H. Moulin et al., 2016), neuroscience (Hsu, Anen, & Quartz, 2008), artificial intelligence (Dickerson, Goldman, Karp, Procaccia, & Sandholm, 2014), justice and policy (Fleurbaey, 2008; Gollwitzer & van Prooijen, 2016; Konow, 2003), and behavioral economics and game theory, especially within the dictator, ultimatum, and estate games (Ashlagi, Karagözoğlu, & Klaus, 2012; Chmura, Kube, Pitz, & Puppe, 2005; Dreber, Fudenberg, & Rand, 2014; Fisman, Kariv, & Markovits, 2007; Huck & Oechssler, 1999; Nowak, Page, & Sigmund, 2000; Pálvölgyi, Peters, & Vermeulen, 2010). Empirical work on fairness is similarly wide-ranging, including work on cultural differences (Gaertner, Jungeilges, & Neck, 2001; Jungeilges & Theisen, 2008; Schäfer, Haun, & Tomasello, 2015; Schokkaert & Devooght, 2003), the developmental trajectory of fairness (Wittig, Jensen, & Tomasello, 2013), the impact of other-regarding preferences in games (Austerweil et al., 2015; Bolton, Brandts, Katok, Ockenfels, & Zwick, 2008), and the weighting of distributive allocation versus the procedure by which items are allocated (Cooney, Gilbert, & Wilson, 2016; Dupuis-Roy & Gosselin, 2011). Wanting quantitative measures, researchers have also developed metrics to quantitatively evaluate the fairness of solutions. Several of these metrics have been compared empirically (Dupuis-Roy & Gosselin, 2009; Engelmann & Strobel, 2004; E. Fehr, Naef, & Schmidt, 2006; Herreiner & Puppe, 2007) and include, for example, envy-free fairness, proportional fairness, and inequality aversion. The tradeoffs between fairness and efficiency (maximizing the joint utility of all agents) are often considered and have been evaluated theoretically (e.g. Bertsimas, Farias, & Trichakis, 2012, 2011).

Much work on fair allocation, however, falls outside our purview, as it considers a self-interested agent deciding how to distribute resources between themself and others. When participants have a personal stake in the situation, biases can appear in their interpretation of what is fair or what behavior they exhibit (Babcock, Loewenstein, Issacharoff, & Camerer, 1995; Beckman, Formby, Smith, & Zheng, 2002; Binmore, 1994; Cappelen, Nielsen, Sørensen, Tungodden, & Tyran, 2013; Croson & Konow, 2009; Ellingsen & Johannesson, 2001; Gächter & Riedl, 2006; Herrero, Moreno-Ternero, & Ponti, 2010; Konow, 2000, 2009; Traub, Seidl, Schmidt, & Levati, 2005). This paper focuses on the perspective of a third party, the scenario in which an impartial decision-maker is choosing how to best help a group of people.

As such, our experiments are most similar to the subset of fair allocation experiments in which the decision-maker's utility is tied only to the utilities of the agents it is serving (e.g. selfless artificial intelligences built to serve human needs). In one such set of studies, participants acted as dictators to fairly allocate goods when the dictator's own payoffs were fixed (Engelmann & Strobel, 2004; E. Fehr et al., 2006). Another two related papers investigated the division of multiple goods by an uninvested agent (Herreiner & Puppe, 2007; Yaari & Bar-Hillel, 1984). In these cases, participants chose how to distribute multiple items across agents whose utility functions were represented in payoff matrices. Our experiments have a similar structure in that they consider preferred allocations over payoff matrices.

Our work is distinct from those previous fair allocation studies, however, in the problem it is solving: here, *the uninvested agent can take one action that has consequences for multiple individuals.* Our problem formulation is more general than the "fair allocation" problem, since the idea of "choosing actions that have implications across multiple utility functions" applies to many situations outside the distribution of items. Specifically, the problem setup we describe, with utilities over any possible action, is more general than having utilities over specific items being distributed. While we are still working with abstract entities in this work, payoff matrices, and in this work only working with positive utilities, this problem formulation admits a range of scenarios. For example, a schoolteacher trying to determine what field trip to bring their students on is an example of our problem, but not a fair allocation problem. Necessarily, our formulation also encompasses the class of fair allocation problems, including problems like sorting multiple indivisible items into "bundles" to be distributed to individuals (e.g. Herreiner & Puppe, 2007). In this case, each "action" is to give each agent each sorted bundle, and if necessary each of these agent-specific actions could be characterized as a single larger action that could be repeated over multiple allocation decisions.

It is also worth emphasizing how the question posed in this paper differs from others in the literature, even those that do not involve a self-interested party. For one, this paper does not singularly employ the zero-sum scenarios present in fair allocation studies. In fair allocation studies, participants distribute a set number of items between agents, and if one agent receives an item another agent loses it. In this work, participants choose to create an item that can bring utility to multiple agents. For these questions, high utility for one agent does not necessarily imply low utility for another.

In addition, this paper asks participants "what *should* be done" rather than what is "fair." Empirically, "fairness" can correspond more to ideas like equality and helping the needy, while questions like "in which society would you like to live?" can evoke responses that take into account the tradeoff between equality and maximizing the benefit to the group. Fairness can also correspond to dis-

tributional fairness or procedural fairness—examining the fairness of the final allocation solutions or the process by which items are allocated. This work focuses not on fairness but on what decisions participants believe a third party should make, in terms of creating a shared item to be distributed.

Similarly, the question of "what should be done" is incredibly dependent on context (see Konow, 2003; Konow and Schwettmann, 2016 for reviews), and our question of what an uninterested artificial intelligence should do constrains the space of those contexts enormously. In many decision-making tasks, arbitrators are contending with preferences for honesty (Dana, Weber, & Kuang, 2007), altruism (Andreoni, 1989; Pelligra & Stanca, 2013), cooperation (Dreber et al., 2014), previous experiences, friendship, spite (Beckman et al., 2002; Levine, 1998), reciprocity (Berg, Dickhaut, & McCabe, 1995; Bolton & Ockenfels, 2000; Cox, 2004; Dufwenberg & Kirchsteiger, 2004), and any number of highly-relevant factors of what is fair and what is preferred. We are interested in what people think should be done in the absence of such considerations: what they would like their leaders or assistive artificial intelligence systems to advise before previously-existing social relationships come into play. Artificial intelligences are by default honest and not more altruistic or cooperative than preferred, so they offer a unique opportunity to act as independent advisors free from spite, obligation, or reciprocity.

Our question could integrate important factors like need or desert / merit (e.g. Alesina & Angeletos, 2005; Fleurbaey, 2008; C. Fong, 2001; Hoffman & Spitzer, 1985; Konow & Schwettmann, 2016; Nord, Richardson, Street, Kuhse, & Singer, 1995; Schokkaert & Devooght, 2003; Schokkaert & Lagrou, 1983; Schokkaert & Overlaet, 1989), but the simplified task we choose does not involve agents who are needy or who have worked harder than others, despite this being a major factor in society. Another way to extend our study would be to use a paradigm that uses bundles or collections of objects, either for allocation or as creating these shared resources (both can be construed as an action). Bundles allow investigation of envy-freeness, and also proportionality, like in "claims problems" when agents may each have a claim to a resource that is greater than the resource is worth (Bosmans & Schokkaert, 2009; Thomson, 2003). Additionally, Konow (2003) discusses the differences between subjective values and objective values, which we simplify here by directly presenting agent utilities. Given the difficulty of the question of what should be done, we used a simplified task and save these questions for future work. For our question and paradigm, we consider it reasonable to limit our literature review to those topics that are most aligned with our question, though we provide references to the reader for additional study.

Our task, then, is to determine what an uninvested, autonomous agent should do in a decision-making scenario with multiple recipients. To investigate what defines a "helpful" action when balancing multiple utilities, we created the following problem setting: we asked what decision a man-

ager should make in choosing a drink for two guests. Participants acted as the manager, repeatedly making choices that we used to develop a model of what people think are good decisions. Drawing on the literature in economics, computer science, and philosophy, we considered four metrics as hypotheses to capture participants' behavior: *maximax*, *maxsum*, *maximin*, and *inequality aversion (IA)*. We determined which combination of metrics best explained participants' behavior by statistical analyses and evaluating the output of a maximum entropy inverse reinforcement learning (MaxEnt) model. Having inferred the preferred metrics, we then used these metrics to compare participants' behavior across conditions.

2.1.1 Metrics

We now delve into the specifics of our metrics and what stimuli they were applied to. Many different fields have investigated the question of how people make decisions when balancing multiple tradeoffs—for example, people could choose to maximize the group utility, or distribute resources evenly. As such, several quantitative models describing "fair" behavior have been suggested. We investigated four metrics that were often used (often in combination) in previous studies to characterize human behavior (e.g. Brams, Edelman, & Fishburn, 2003; Dupuis-Roy & Gosselin, 2009; Engelmann & Strobel, 2004; Herreiner & Puppe, 2007). We do not believe these metrics span the space of reasonable preferences people may hold—people's algorithms for determining what to do in the world can be extremely complex—but we selected these metrics based on what has been widely and empirically observed in the literature. We applied these four metrics to a set of payoff matrices $\mathcal{M}$: in each matrix $\mathcal{M}_m$, two agents' utilities (A and B) were shown for each of four drink options (see Fig. 2.1). We investigated each of the metrics for these payoff matrices.

Intuitively, one method to select a shared option is to add up the utilities of all agents, and pick the option that maximizes this joint utility value. This is a metric called *maxsum*, which maximizes the happiness of the group rather than individuals. It is a Pareto optimal option. In Fig. 2.1, the best *maxsum* option would be the third cup, because when utilities of agents A and B are added together, these sums are: 7 (first cup), 13 (second cup), 15 (third cup), 14 (fourth cup). A different metric enforcing fairness is *maximin*: maximizing the utility of the person who is worst-off. Intuitively, this metric means making sure that no individual agent is very unhappy. In Fig. 2.1, the best *maximin* option would be the second cup, because when we look to which of the agents are worst-off in each of the pairs, these agents' utilities are: 3 (first cup), 5 (second cup), 4 (third cup), and 2 (fourth cup), and the second cup leaves the worst-off agent with the highest possible utility. Another measure of fairness is *inequality aversion (IA)*: decreasing the difference in utilities between

agents. Intuitively, this metric means that agents should be equally happy. In Fig. 2.1, the best *IA* option would be the first cup, because the agents' utilities will be maximally close to each other. A final possible measure is *maximax*: maximizing the highest utility, searching across both agents. Intuitively, this metric means that any one agent should be made as happy as possible. This option might not initially seem fair, but if presented with the opportunity for repeated choices, pleasing one agent each round may seem like the best solution. In Fig. 2.1, the best *maximax* option would be the fourth cup, because this cup allows one agent to have its highest possible utility. Note that in Fig. 2.1, the best example of each metric was a different cup, but in most of our payoff matrices, the best example of multiple metrics was the same cup. Thus, in this paper we evaluated the four metrics of *maximax*, *maxsum*, *maximin*, and *inequality aversion (IA)*.

Metrics like *maximin* and *maxsum* have a long history in the literature. The concept of *maximin* was popularized by Rawls (1971, 1974), who advocated for allocating resources to the least well-off individual. Harsanyi (1975) argued in favor of expected utilitarianism, loosely the *maxsum* principle, except in cases where the *maximin* choice was similar to the *maxsum* choice, as in the case in our experiments. There has been a strong focus on *maximin* mathematically and in economics (e.g. Amanatidis, Markakis, Nikzad, & Saberi, 2017; Barman & Krishna Murthy, 2017; Dubois, Fargier, & Prade, 1996; Escoffier, Gourvès, & Monnot, 2013; Kurokawa, Procaccia, & Wang, 2016; Procaccia & Wang, 2014), including applications to networking (e.g. bandwidth-sharing) (Salles & Barria, 2008). Choices aligned with the *maximin*, *maxsum* metric, and *inequality aversion* metrics are often compared in studies, and results are often mixed, with participants trading off between different metrics depending on the numbers and contexts involved (Ahlert, Funke, & Schwettmann, 2013; Charness & Rabin, 2002; Engelmann & Strobel, 2004; Faravelli, 2007; E. Fehr et al., 2006; Gaertner et al., 2001; Gaertner & Schwettmann, 2007; Konow, 2001, 2003; Konow & Schwettmann, 2016; Mitchell, Tetlock, Mellers, & Ordonez, 1993; Ordóñez & Mellers, 1993; Pelligra & Stanca, 2013; Schwettmann, 2009, 2012). These studies differ in their experimental paradigms, and results have been subsequently different, even as the concept of tradeoffs between a few principles of justice have remained similar (Konow, 2003). These experiments reveal a few common difficulties as well. Many experiments attempt to target each metric in isolation, which does not account for the correlations between choices: a choice that maximizes the *maximin* metric also tends to rate highly on the *inequality aversion* metric and less well on the *maxsum* or *maximax* metrics. Additionally, experiments often present somewhat extreme choices and models contain variables that are occasionally confounded. We aim to address these difficulties in our study by presenting many nuanced choices to participants, and then using a computational model and statistical analyses to account for correlations and confounding between metrics. Using these tools, we gain the ability to distinguish the

relative influence of metrics on sets of participant choices.

In this paper, we aim to address the question of what policies people would like decision-makers, and the assistive technologies assisting decision-makers, to have in the future. To this end, we focus on research studies that are aimed at non-interested third parties reasoning about helpful decision-making, the results of which are not always consistent. In these studies, participants are presented with options that implement various metrics, and have to choose how to allocate items across agents (Engelmann & Strobel, 2004; E. Fehr et al., 2006; Herreiner & Puppe, 2007; Yaari & Bar-Hillel, 1984). Participants are often shown choices that are used to disambiguate between metrics, and because these choices are so distinct, participants may only make a single or small number of choices for any given prompt. We took a different approach with our work: on each prompt we presented many choices to participants, accepting the high amount of overlap among metrics necessary to describe participants' responses. We constructed a computational model to accommodate these correlations, and used this model to reveal the relative contributions of different metrics across many probes of participant behavior. We tested the generalizability of our findings by ensuring the participant behavior was similar across different prompts. Additionally, previous work often focuses on short-term, single decisions; we investigated participants' intuitions in the repeated setting where they could make more long-term decisions. In summary, our work compares four common metrics in a setting where participants are making more choices than in previous work and the correlations among metrics describing those choices are accounted for. We use this higher resolution into the relative contributions of each of these metrics to directly compare them, and probe many questions—including longer-term decision-making—to determine the generalizability of our findings. These improvements take place in a novel setting, in which participants are not being asked about how to allocate many items, but to create a resource to be shared among multiple agents. We thus present a novel test of how correlated metrics interact, in the context of how people feel third-party decision-makers should balance others' utilities: a problem formulation that will occur more and more often in technologies in the future.

In three experiments, we investigated the contributions of the metrics of *maximax*, *maxsum*, *maximin*, and *inequality aversion* in describing what people consider good or helpful behavior. In Experiment 1, we examined participants' choices in the drink task and determined whether these choices were consistent when the hypothetical decision-maker was described as a human or a robot, and when the agents receiving the resources were described as friends or strangers. Our results indicate that participants were consistent in using the *maximin* metric to make decisions (maximizing the utility of the worst-off agent) despite variations in the agents involved. In Experiment 2, we asked what decisions people would make when they had the opportunity to offer more than one

drink to the same set of agents. We tested whether people would make choices while considering the entire multi-decision expected utility, or focus on the individual decisions within. We found that participants reasoned over the entire multi-decision process, and the *maximin* metric could again describe their choices. In Experiment 3, we validated our results from Experiment 2 by presenting participants with the cumulative sum of the choices available in Experiment 2, and observing that when the multi-decision problem was condensed to a single instance again, participants reliably made choices described by the *maximin* metric.

Figure 2.1: Example matrix $\mathcal{M}_m$. Columns are options o^j, while rows show each agent i's utility $\mathcal{U}^i$. Here, the problem is to decide which drink to serve to both agents A and B.

2.2 Experiment 1: Which metrics describe people's behavior? Are they robust to changes in agent?

Here we tested which combination of metrics described participants' behavior on a drink-choosing task (Fig. 2.1). We tested a variety of different phrasings and altered scenarios to ensure that the results were consistent across changes in presentation. Specifically, we tested whether participants' choices changed depending on if the decision-maker was an artificial intelligence (a robot) or a human, whether the recipients of the drinks were friends or strangers, and whether the participant was stated to be one of the recipients of a drink. We might expect that a participant would perform differently in the "Robot" condition if they thought that what a robot should do in a manager role was

different from what a human manager should do. For example, participants may think that artificial intelligences need to remain completely impartial or compute everything exactly, whereas humans should rely on gut instincts. Similarly, we might expect that participants would have different intuitions about a server giving drinks to strangers rather than friends. Perhaps participants would think that when serving friends, a server should worry more about joint happiness rather than making sure utilities were equitable, since friends could make it up to each other later, while the same wouldn't be true of strangers. We were also interested in whether participants' opinions of what decision should be made would change if they themselves were receiving an item rather than hypothetical recipients. In Experiment 1 we were aiming to test whether participants' empirical intuitions of good decision making would extend across these variations in scenarios, which is important for the generalizability of our findings.

2.2.1 Method

Participants

Participants with U.S. IP addresses were recruited from Amazon Mechanical Turk across five conditions: "Nominal" (n = 36, 0 participants excluded), "Robot" (n = 35, 1 participant excluded), "Robot Friends" (n = 35, 1 participant excluded), "Robot Strangers" (n = 34, 2 participants excluded), and "Veil of Ignorance (VoI)" (n = 33, 3 participants excluded). Participants were paid between \$2.50-\$3.00 for their participation. Participants were excluded if they failed the included attention check or indicated that they did not understand the experiment.

Stimuli

Stimuli were chosen such that participants would reason over a set of positive-utility actions, and the effects of each metric could be quantitatively isolated from this data. To keep the paradigm simple, utilities in the form of numbers in a table were used ("payoff matrices"), with utilities kept small to allow participants to use simple addition. To maximize the generalizability of the findings, several constraints on the matrices were put in place to ensure participants were reasoning over a wide, randomized range of independent matrices.

Participants viewed 20 of these payoff matrices (set of matrices $\mathcal{M}$), one at a time. Each matrix $\mathcal{M}_m$ was 2×4, where the rows indicated the two agents, the columns indicated the four colored drinks, and each agent's utility for each drink was shown.

Matrices were generated by rejection sampling. Matrices were subject to the following general constraints. The sum of agents' utilities for each option o^j, $\sum_i \mathcal{U}^i(o^j)$, was constrained to be within

2 and 16, so that there would be a wide range of choices available but participants would only need to use simple addition. Within each matrix, columns (both agents' utilities for an option) were not allowed to repeat, including permutations within columns, so that there would always be four independent choices within a matrix. Moreover, within each matrix, for every column, there could be no other column that strictly dominated that column according to all metrics, because such a column would rarely be picked and so would be uninformative according to the goals of this study. Since we were interested in the question of what desired resource should be shared among recipients, we constrained our actions to positive utilities and did not allow matrices to contain zeros or negative numbers. Finally, in a set of matrices $\mathcal{M}$, any two matrices were not permitted to have more than two of the same columns, where "sameness" included column permutation, in order to create a set of independent matrices.

In addition to the general constraints, "class" constraints were established to ensure that the chosen matrices could isolate the effects of each of the metrics. There were thus four classes: *1*) 4 matrices in which each option was the best choice according to one of the metrics (e.g. the example in Fig. 2.1), *2*) 4 matrices in which the *maxsum* metric was held constant: all choices had the same joint utility, *3*) 4 matrices in which the *maximin* metric was held constant: for all choices, the worst-off person had the same utility, and *4*) 8 matrices that were randomly generated. The *maximax* and *inequality aversion* metrics were not held constant because matrices constructed in this manner did not meet the general constraints described above.

Procedure

Upon viewing each matrix, participants read the following text: "You're the manager at a hotel and want to serve a drink to the room. Archie and Ben are your guests and have told you how much they enjoy different drinks (higher numbers mean more enjoyment). Which drink would you like to serve?" Participants then had to select one of the four drinks, and justify their response. The names "Archie" and "Ben" were substituted with other names beginning with "A" and "B" for each matrix.

We expected that participants would employ a consistent understanding of what made a "good" decision in the drink-choosing task, but wanted to test the robustness of participants' choices by employing several task variations. The variations we tested were the identity of the server (either human or robot), the relationships of the agents being served (either friends or strangers), and whether the participant was described as a recipient of a drink.

In the "Nominal" condition, participants were presented with the default text ("You're the manager at a hotel and want to serve a drink to the room..."). In the "Robot" condition, the prompt

was: "There is a robot manager at a hotel which will serve a drink to the room. Archie and Ben are its guests and have told it how much they enjoy different drinks (higher numbers mean more enjoyment). Which drink would you like the robot to serve?" In the "Robot Friends" condition, the prompt was the same as in the "Robot" condition, but the following phrase was added: "Archie and Ben are its guests *(they are friends with each other)*...". In the "Robot Strangers" condition, the following phrase was substituted: "Archie and Ben are its guests *(they are strangers to each other)*...". Italics were not included in the participant text. In the "Veil of Ignorance (VOI)" condition, the instructions were modified to be: "The manager at a hotel wants to serve a drink to the room, where you and another guest are sitting. The manager has learned how much you both enjoy different drinks (higher numbers mean more enjoyment). Given you don't know which guest (A or B) you are, which drink would you like the manager to serve?" This condition is named the "Veil of Ignorance (VOI)" condition as a reference to Rawls (1971), who suggested that a method of eliciting moral judgments without self-interest would be to present scenarios in which participants had to make judgments about hypothetical societies they would like to live in, while not knowing anything about their place in the hypothesized social order. Thus, participants would be "behind a veil of ignorance" in that they would have to make decisions without knowing whose place they would fill. Here, in this experimental condition, participants were told they were recipients of a drink but did not know which recipient (A or B) they were, thus enacting a version of the Rawlsian thought experiment.

To analyze participant responses according to our metrics, we calculated the value of each metric ($\mathcal{F}_q$), where q indexes the individual metric (*maximax, maxsum, maximin, IA*), and $\mathcal{F}$ is a vector. Values for the metrics were calculated from the utilities of agent i ($\mathcal{U}^i$) for each option o^j:

$$\mathcal{F}_{maximax}(o^j) = \max_i \mathcal{U}^i(o^j)$$
$$\mathcal{F}_{maxsum}(o^j) = \sum_i \mathcal{U}^i(o^j)$$
$$\mathcal{F}_{maximin}(o^j) = \min_i \mathcal{U}^i(o^j)$$
$$\mathcal{F}_{IA}(o^j) = \prod_i \frac{\mathcal{U}^i(o^j)}{\sum_i \mathcal{U}^i(o^j)}$$

These $\mathcal{F}_q(o^j)$ values were then normalized to account for the tradeoffs between each option o^j in the matrix $\mathcal{M}_m$[1] :

$$\mathcal{F}_q(o^k) = \frac{\mathcal{F}_q(o^k)}{max_{o^j \in \mathcal{M}_m}\{\mathcal{F}_q(o^j)\}}, \quad \forall q \tag{2.1}$$

[1]The disadvantage of normalizing the values within each metric was that information about metrics that had particularly high or low values for a given matrix was lost. However, we considered this choice better than the alternative unnormalized option, for which there would be no sense of the alternative options participants were choosing between.

2.2.2 Results and Discussion

We were interested in what metrics participants preferred, meaning which metrics could describe participants' demonstrated behavior. We determined preferred metric through an independent statistical analysis, and then via a maximum entropy model. We examined the proportion of participants preferring specific metrics in the "Nominal," "Robot," "Robot Friends," "Robot Strangers," and "Veil of Ignorance" conditions.

Statistical analysis

We sought to determine which metrics individuals used to make their choices. Before asking which metrics *best* described participants' choices, we first asked whether participants were behaving according to any of the metrics, by comparing participants' empirical choices to a simulated baseline participant making random choices. For each metric q, we determined whether the $\mathcal{F}_q$ values from participants' empirical choices were significantly larger than the $\mathcal{F}_q$ values from random choices, summed across participants and matrices. For our baseline comparison, we drew from the null distribution to generate enough choices for a complete experiment (choices for each matrix and for each participant, summed), and repeated that process 10000 times. We then computed z-scores and associated p-values for whether the empirical $\mathcal{F}_q$ values (H1) were significantly greater than our baseline $\mathcal{F}_q$ values (H0) for each metric q. We found that the *maxsum*, *maximin*, and *IA* metrics significantly matched participant behavior, and that the *maximin* metric best matched participant behavior (had the highest z-scores) for all conditions except the "Veil of Ignorance" condition (Table 2.1). In the "Veil of Ignorance" condition, the *maxsum*, *maximin*, and *maximax* metrics significantly matched participant behavior, and the highest z-score was for the *maxsum* metric, closely followed by the *maximin* metric. This analysis was chosen because the comparison between the empirical and null distributions accounted for the biases within metrics, and choice of specific matrices.

Inferring metrics used with a maximum entropy model

To further test which combination of metrics described participants' choices, we constructed a maximum entropy inverse reinforcement learning (MaxEnt) model (Ziebart, Maas, Bagnell, & Dey, 2008). In this model, we inferred weights, which were associated with each of the metrics. Mathematically, the weight vector θ was indexed by q and composed of: $[\theta_{maximax}, \theta_{maxsum}, \theta_{maximin}, \theta_{IA}]$. Participants' preferred metric was evaluated as the metric with the largest associated weight θ_q.

We computed a maximum *a posteriori* estimate for θ for each participant given their choices by combining a uniform prior over the parameters of θ with the likelihood of each individual's choice

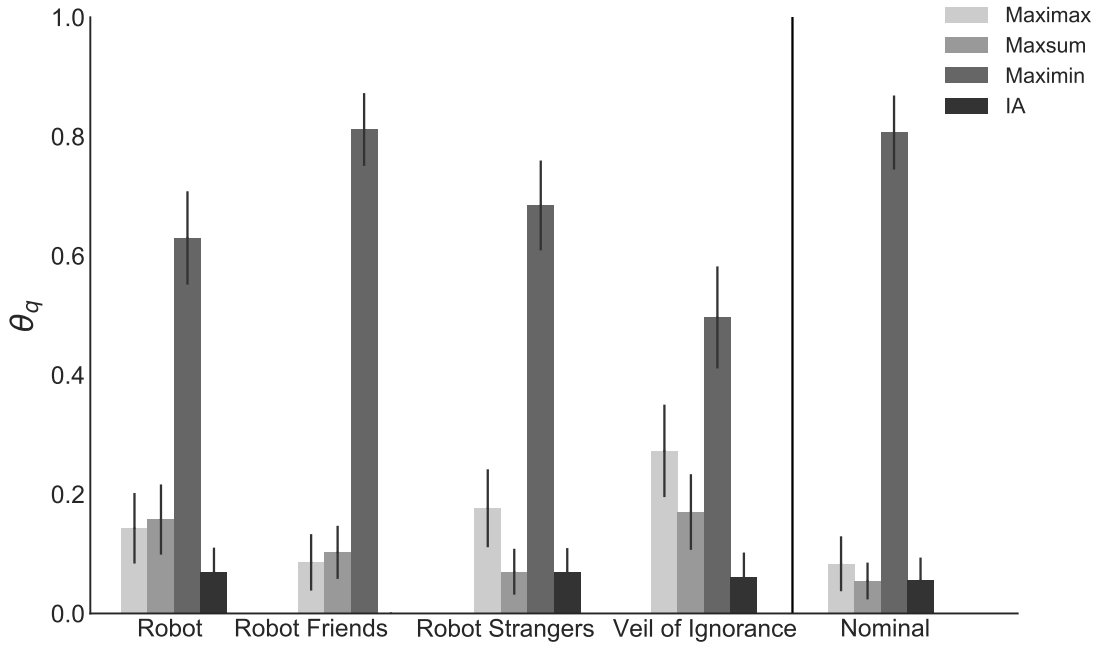


Figure 2.2: MaxEnt model results for Experiment 1, showing mean inferred weight $\theta \pm$ standard error for each metric across participants in the "Nominal," "Robot," "Robot Friends," "Robot Strangers," and "Veil of Ignorance" conditions. The *maximin* metric best described participant behavior.

o^k over the options o for each matrix. Specifically, within each matrix $\mathcal{M}_m$, for each participant's choice o^k over options o, we maximized the function $\log P(o^k \mid \theta)$ over θ (a vector of length q), where:

$$P(o^k \mid \theta) = \frac{e^{\theta^\top \mathcal{F}(o^k)}}{\sum_j e^{\theta^\top \mathcal{F}(o^j)}} \tag{2.2}$$

Recall that $\mathcal{F}$ is the vector containing the values of the individual metrics $[\mathcal{F}_{maximax}, \mathcal{F}_{maxsum}, \mathcal{F}_{maximin}, \mathcal{F}_{IA}]$.

We summed across matrices $\mathcal{M}_m$ to create the final cost function $\sum_m log\ P(o^k(\mathcal{M}_m) \mid \theta)$, and used the following constraints to compare the relative use of each of the metrics: $\sum_q \theta_q = 1, \theta_q \geq 0$. We optimized using sequential least squares programming (SLSQP) in Python's *scipy* package. To calculate the weight vector of an average individual, we calculated the arithmetic mean over individual participants' θ vectors. Thus, overall the θ_q value for any given metric was determined by the closeness between the responses expected under that metric (e.g. *maxsum*) and a participant's responses. In interpreting the results, if any given θ_q value was high, then participants often chose responses in accordance with that metric.

We found that participants' behavior was best explained by the *maximin* metric for all conditions according to the MaxEnt model, as shown in Fig. 2.2. Our conclusion that participants' behavior was best explained by the *maximin* metric was also supported by using an alternative method of calculating the average weights across participants. Throughout this work, we sought to determine which combination of metrics described an "average" individual; however, there are several ways of calculating an average weight vector. In the main analysis, we set "average" as the arithmetic mean of each individual participant's weight vector θ. An alternative average assumes that all participants' data came from one individual, and given this, calculates a single weight vector θ under the assumptions of the MaxEnt model described above. Using this, $\theta_{maximin} = 1$ and $\theta_{maximax}, \theta_{maxsum}, \theta_{IA} = 0$ for the "Nominal," "Robot," "Robot Friends," and "Robot Strangers" conditions.

Note that the results from the MaxEnt model support and also further the results from the statistical analysis. In the statistical analysis for Table 2.1, we compared participants' responses to those expected from a participant choosing randomly. We saw, compared to this null distribution, that participants' choices tended to encompass the *maxsum*, *maximin*, and *inequality aversion* metrics across the "Nominal," "Robot," "Robot Friends" and "Robot Strangers" conditions. While the results from that analysis suggested that participants' responses slightly more matched the responses expected from a *maximin* policy (indicated by the higher z-scores under the *Maximin* comparison) compared to other metrics' policies, this comparison was indirect (participant and random responses were compared along each of the metrics, and then the metrics were compared, rather than directly comparing the relative influence of each metric in participant responses). To more directly compare which metrics best described participants responses, we examine the results from the MaxEnt model designed for this purpose. The results from the MaxEnt model for these conditions clearly differentiate the *maximin* metric as more descriptive of participants' responses compared to the other metrics.

For the "Veil of Ignorance" condition, the MaxEnt results were more evenly split between the *maximin* and *maxsum* metrics: $\theta_{maximin} = .52$, $\theta_{maxsum} = .48$, and $\theta_{maximax}, \theta_{IA} = 0$. Of note, while the MaxEnt results emphasized the *maximin* and *maximax* results in Fig. 2.2, the results from fitting a single θ emphasize the *maximin* and *maxsum* results. This pattern could arise if there were some participants who behaved very consistently according to the *maximax* metric, leading to their being highly represented when θ values were calculated for each participant and averaged, but these participants' behavior washing out when all participants' behavior was aggregated. It is noteworthy that the *maximax* and *maxsum* metrics were highly correlated (the same is true for the *maximin* and *IA* metrics) such that if participants were commonly making the best *maximax* choices, these

choices were likely to be well-described by the *maxsum* metric as well (Fig. 2.3, leftmost). This correlation is how there could be a result for the single θ providing more support for the *maxsum* metric with an average mean providing more support for the *maximax* metric. The shared emphasis on the *maximin*, *maxsum*, and *maximax* metrics in the "Veil of Ignorance" condition is also evident from the statistical analysis described above. In sum, across each of these result measures behavior of participants in the "Veil of Ignorance" condition appears to be described by the *maximin*, *maxsum*, and *maximax* metrics.

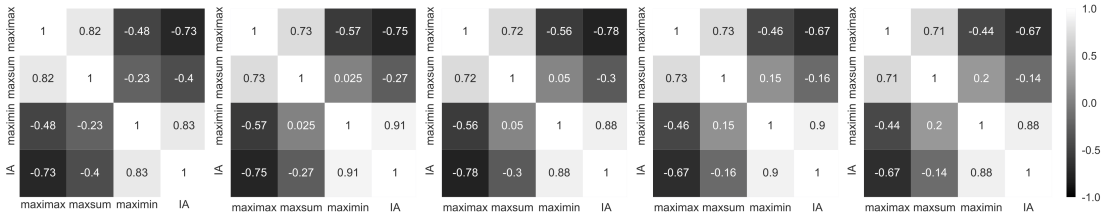


Figure 2.3: Correlations across values of the metrics in the "Nominal" (**leftmost**), "Follow-Up (2x)" (**left**), "Follow-Up (3x)" (**center**), "Repeated 2x: Summed" (**right**), and "Repeated 3x: Summed" (**rightmost**) conditions. For each condition, we computed the value of each metric for all possible choices across all matrices. We concatenated all of the possible choices across all matrices into one vector for each metric, and computed the Pearson correlation coefficient across metrics. Choices well-described by the *maximax* metric were also well-described by the *maxsum* metric (high correlation between *maximax* and *maxsum*), while choices well-described by the *maximin* metric were also well-described by the *inequality aversion (IA)* metric (high correlation between *maximin* and *IA*). These clusters (*maximax* and *maxsum*) and (*maximin* and *IA*) tended to be anti-correlated. Note that the "Nominal" condition plot also describes the "Robot," "Robot Friends," "Robot Strangers," "Veil of Ignorance," "Repeated 2x: Independent," and "Repeated 3x: Independent" conditions, since participants saw the same stimuli. The summed possible choices used in the "Repeated 2x: Summed" and "Repeated 3x: Summed" correlation matrices were used only for data analysis and not shown to the participants.

The MaxEnt model results were designed to isolate which metric best described participants' behavior by accumulating evidence over all trials. As such, the MaxEnt model seems to clearly indicate participants were always behaving according to, for example, the *maximin* metric. However, in each individual trial, participants could not behave according to an isolated metric, since each choice they made provided some evidence for each of the metrics. This inseparability—caused by the correlations among metrics, see Fig. 2.3—is what motivated our use of the MaxEnt model. One could argue that participants may be making choices that were most supporting a single metric on each trial, and this is possible but not what was empirically observed. Though we do not include trial-by-trial data, Table 2.2 provides a histogram of the metrics participants' behavior most adhered to, according to the MaxEnt model. Participant behavior generally was in accordance with several of

the metrics, and most in accordance with the *maximin* metric in aggregate.

After constructing the MaxEnt model, we wanted to check that our weight vectors θ generalized and that the model was predictive of human behavior. To that end, we estimated participants' weight vectors and used these to predict held-out choices. Specifically, given participants' weight vectors θ, we could predict participants' choices given a new set of options in matrix $\mathcal{M}_m$. The predictions o_m^k were calculated by $\arg\max_j P(o_m^j \mid \theta)$. We calculated weight vectors θ for 50% of participants, each trained on 50% of the matrices. We then used the average weight vector from these participants to predict the remaining 50% of participants' choices on the remaining 50% of matrices, and report this prediction accuracy. As a comparison, we also report the accuracy of prediction when each participant's weight vector, trained on all matrices, is used to predict that same participant's choices on each matrix. We report the average of 100 training runs for each prediction.

Training and testing on 50% participants and matrices, the "Nominal" condition had 72.9% predictive accuracy (non-held out comparison: 75.4%); the "Robot" condition had 66.4% predictive accuracy (non-held out comparison: 70.1%); the "Robot Friends" condition had 71.3% predictive accuracy (non-held out comparison: 70.9%); the "Robot Strangers" condition had 64.6% predictive accuracy (non-held out comparison: 73.3%); the "Veil of Ignorance" condition had 59.5% predictive accuracy (non-held out comparison: 78.2%). Chance accuracy was 25%. We observed high predictive accuracy both when testing sets were and were not held out compared to chance accuracy, supporting the validity of our inferred θ vectors.

What metric was preferred across conditions?

We also asked whether participants' judgments of fairness would differ based on what type of agent they were considering, and whether the participant was considered a drink recipient. We found that on the whole, participants were consistent in their behavior across conditions: whether they were thinking about a human manager or a robot manager, whether the robot manager was serving friends or strangers, and whether they were to receive a drink. Specifically, participants' preferred metric did not differ significantly between the "Nominal," "Robot," "Robot Friends," "Robot Strangers," and "Veil of Ignorance" conditions (χ-squared test over $\mathcal{F}_q$ summed over participants and matrices: $\chi^2 = 6.67$, $p = .88$, $df = 12$). We could have generated many more conditions, but our results suggest that participants' fairness judgments are robust to at least some changes in agent identity, and generally that participants' ideas of fairness are similar across varied situations.

The result of similar behavior across conditions is important, because in this work our goal is to examine humans' views of what decisions decision-makers and assistive artificial intelligences should

implement in the future. In this experiment we asked participants what a desirable decision would be, and then checked that the results were not a specific consequence of slight differences in how we could have asked the question. There were indeed no significant differences in participants' reports of what they considered a good decision, despite differences in the agent considered. This finding is important with respect to the generalizability of our findings and how we might outsource decisions to technology in the future.

While results were similar across conditions, it is worth noting that the results from the "Veil of Ignorance" condition subtly differed from the rest, for example by encouraging behavior more closely matching the *maxsum* metric. In the "Veil of Ignorance" condition, participants seemed to engage with the question as a new context: qualitatively, participants seemed to think less about fairness and more in the sense of "gifts" or "gambling." When asked to justify their choices, participants in the "Veil of Ignorance" condition would say things like: "one of us might as well be very happy," "I don't know what guest I am so I made the safest choice," or "neither of us will get something we don't enjoy at all." Whereas in the other conditions, participants tended to more often use words like "fair," "balanced," or "equal." A possible explanation for why behavior in the "Veil of Ignorance" condition was not so closely matched to the *maximin* metric as was true in the "Nominal," "Robot," "Robot Friends," and "Robot Strangers" conditions is this apparent difference in framing. On the one hand (corresponding to the "Veil of Ignorance" condition) the participant could either gamble to win a desired item or offer it generously to an opponent, or alternatively play it safe with a drink everyone would like a little (Konow and Schwettmann (2016) reviews the evidence on whether "fair" behavior is actually risk-averse behavior). On the other hand (corresponding to the other conditions), the participant would be responsible for the outcomes of two people one doesn't know, perhaps encouraging more "fair" allocation strategies.

While participants' behavior was still more *maximin*-aligned than *maxsum*-aligned, the fact that participants' behavior was relatively more similar to the *maxsum* metric in the Veil of Ignorance condition was interesting because the participants could have taken the opposite perspective. In the Veil of Ignorance condition, the participants had an (unknown) stake in the proceedings and could have made sure not to end up with the unfair end of a bargain by engaging in even more *maximin* behavior relative to the other conditions. Instead, participants behaved relatively more according to the *maxsum* metric. This result is important and should be further explored, since most of us are recipients and not the decision-maker in many real-life situations, perhaps making the Veil of Ignorance condition the most descriptive of real situations.

Various other studies have empirically evaluated preferences under the veil of ignorance (Andersson & Lyttkens, 1999; Carlsson, Gupta, & Johansson-Stenman, 2003; Frohlich, Oppenheimer,

& Eavey, 1987; Johansson-Stenman, Carlsson, & Daruvala, 2002; Oleson, 2001). Of the studies that evaluated metrics similar to ours, Frohlich et al. (1987) found that participants preferred maximizing average income with a floor constraint, and Oleson (2001) found that participants demonstrated both *maximin* and *inequality aversion* behavior—however, this collection of studies was different enough from our paradigm (e.g. studying risk aversion, or looking at welfare over an entire hypothetical society) that the different contexts likely significantly affected outcomes. A study by Bosmans and Schokkaert (2004) was similar to our work in that they asked participants about their preferences directly, and also under a veil of ignorance. They observed different results between the directly-elicited preferences and veil of ignorance conditions (but not maximally different results, which occurred in comparing directly-elicited preferences to a third self-interested condition). These results were similar to ours, as we found a small difference in response profiles between our "Nominal" and "Veil of Ignorance" conditions. A final aspect of our "Veil of Ignorance" condition is that since our question was about autonomous third-party decision-makers with no stake in the game, we designed the condition such that participants would not receive a payout according to their choices. If participants had a stake in the game (albeit an unknown one), their behavior might have shifted to be more conservative under the assumption of maximizing self-interest—it is known that participants do change their behavior based on whether they are being paid or not (e.g. Gächter & Riedl, 2006; Herrero et al., 2010). Altogether, the "Veil of Ignorance" results were similar to those of the other conditions, but future work will have to continue evaluating the contexts that influence participants' conceptions of good decision making.

	Maximax	*Maxsum*	*Maximin*	*IA*
Nominal	-.2 (.58)	11.2(0)	19.7(0)	10.9(0)
Robot	.6(.27)	10.7(0)	16.7(0)	8.6(0)
Robot Friends	1.6(.05)	13.6(0)	19.0(0)	9.7(0)
Robot Strangers	.6(.27)	9.5(0)	14.7(0)	7.3(0)
VoI	3.1 (.002)	8.7(0)	7.8(0)	-.1(.53)
Rep2x: Ind.(C1)	4.1(0)	9.7(0)	7.8(0)	1.9(.03)
Rep2x: Ind.(C2)	1.3(.10)	5.6(0)	6.9(0)	2.8(.002)
Rep3x: Ind.(C1)	-1.2(.89)	5.7(0)	11.9(0)	6.9(0)
Rep3x: Ind.(C2)	.4(.34)	3.6(1e-4)	5.4(0)	3.0(.001)
Rep3x: Ind.(C3)	-2.1(.98)	1.8(.03)	7.3(0)	5.0(0)
Rep2x: Summed	-9.4(1)	9.4(0)	19.2(0)	14.4(0)
Rep3x: Summed	-24.0(1)	-.007(.50)	26.1(0)	20.8(0)
Follow-Up (2x)	-11.0(1)	5.3(0)	22.1(0)	14.7(0)
Follow-Up (3x)	-6.4(1)	7.8(0)	18.8(0)	9.9(0)

Table 2.1: Results for Experiment 1, showing relationship between participants' responses and hypothetical null distribution for each metric. Determining which metrics describe participant choices, as compared to what would be expected given random choices, summed across participants and matrices. Z-scores are shown, calculated as the [empirical (H1) summed scores − mean of the null (H0) distribution summed scores] divided by [standard error for the null (H0) distribution summed scores], along with associated p-values in parentheses, for all metrics. (Summation is across all matrices and participants.) High scores for any metric indicate that participants often chose responses in accordance with that metric, above and beyond what they would have done had they been choosing randomly. Results are shown for all conditions in all experiments: "Nominal," "Robot," "Robot Friends," "Robot Strangers," "Veil of Ignorance," "Repeated 2x: Independent (Choice 1)," "Repeated 2x: Independent (Choice 2)," "Repeated 3x: Independent (Choice 1)," "Repeated 3x: Independent (Choice 2)," "Repeated 3x: Independent (Choice 3)," "Repeated 2x: Summed," "Repeated 3x: Summed," "Follow-Up (2x)," and "Follow-Up (3x)." 10000 samples of summed scores from the null distribution were taken. Condition names are shortened: "VoI" represents the "Veil of Ignorance" condition and "Rep2x: Ind.(C1)" represents the "Repeated 2x: Independent (Choice 1)" condition. "0" in the table refers to $<.0001$.

	Maximax	*Maxsum*	*Maximin*	*IA*
Nominal	8	3	83	6
Robot	14	17	63	6
Robot Friends	9	9	83	0
Robot Strangers	18	6	71	6
VoI	27	18	48	6
Rep2x: Ind.(C1)	29	12	58	0
Rep2x: Ind.(C2)	25	21	46	8
Rep3x: Ind.(C1)	14	10	62	14
Rep3x: Ind.(C2)	19	29	38	14
Rep3x: Ind.(C3)	29	5	48	19
Rep2x: Summed	0	12	87	0
Rep3x: Summed	10	5	71	14
Follow-Up (2x)	3	6	90	0
Follow-Up (3x)	17	0	83	0

Table 2.2: Percentage of participants whose behavior is best described by each metric, according to the MaxEnt model. Each participant's final θ vector according to the MaxEnt metric was computed for each condition; the highest-value θ_q value was taken as the metric that best described the participant's behavior. Shown is the percentage of participants for which each metric best described their behavior. These more individual-level results closely resemble the aggregated results shown in the Figures. Condition names are shortened: "VoI" represents the "Veil of Ignorance" condition and "Rep2x: Ind.(C1)" represents the "Repeated 2x: Independent (Choice 1)" condition.

2.3 Experiment 2: Repeated choices

Previously we evaluated what people thought a decision-maker should do when taking a single action. However, in the real world, decision-makers will act many times: the schoolteacher choosing a new lesson plan for their students every day, or the government delivering many aid programs over time. Here we investigated whether people have different intuitions for what decision-makers should do when facing repeated decisions with the same set of agents. Perhaps the decision-maker should give one agent its best choice the first time, and a different agent its best choice the second; or perhaps the conclusion is to compromise each time, or to attempt to maximize fairness across the sum of all decisions. We tested whether different metrics would describe participants' actions when participants were given repeated choices.

Repeated decision-making has been studied in the literature, often under the name of sequential or dynamic choices in game theory games. In the repeated version of the Prisoner's Dilemma, participants choose whether to betray their partner or cooperate on each turn, and there has long been research into the optimal strategy for this game (Andreoni, Miller, et al., 1993; Boyd & Lorberbaum, 1987) and other competitive/cooperative games (Kuzmics, Palfrey, & Rogers, 2014), wherein each partner obtains more information about the other on each turn. Behavior in repeated games is sensitive to anticipated repetitions: game outcomes are empirically different when participants are engaging in a finitely repeated game compared to a repeated one-shot game (Andreoni, Miller, et al., 1993). In other sequential games, the sequence is not of both agents making choices simultaneously and then repeating the game, but rather of agents taking turns right after each other, for example by claiming an item via sequential allocation (Bouveret & Lang, 2011; Kalinowski, Narodytska, & Walsh, 2013), or fair queueing (Demers, Keshav, & Shenker, 1989; Hervé Moulin & Stong, 2002). There are also "online" games, in which agents may arrive at different times (e.g. Kash, Procaccia, & Shah, 2014; Walsh, 2011) and resources must be divided between them repeatedly. Alternatively, items may arrive over time to the same set of agents who have preferences over them (e.g. Aleksandrov, Aziz, Gaspers, & Walsh, 2015), which is more similar to the structure of our experiment.

There is a subset of research studying the situation of a single item being distributed to a set of agents repeatedly. In this case, researchers are often testing optimum strategies for allocation that maximize total agent utility, while ensuring that their strategy has useful qualities like being efficient and fair, and encouraging truth-telling of preferences from each agent on each round. This work can fall under the heading of "dynamic mechanism design," and has been widely investigated (e.g. Bergemann & Valimaki, 2006; Cavallo, 2008; Guo, Conitzer, & Reeves, 2009). Other researchers have argued that these paradigms do not capture the structure of real-world problems, and so in-

troduced a real-life food distribution problem in which a central decision maker must repeatedly allocate food to different charities in an online fashion over complex preferences where fairness and efficiency are both considered (Aleksandrov et al., 2015; Walsh, 2015). Our Experiment 2 is similar to this food distribution problem in that we have a repeated task in which a central decision maker makes resource choices over the preferences of several agents, and similar to the dynamic mechanism design studies in that there is a single item being repeatedly allocated. Our Experiment 2 is different, however, in that the single item is not being allocated to a single recipient, but rather being created and shared across agents on every round.

Freeman, Zahedi, and Conitzer (2017) have perhaps the most similar motivation to our experiment, as they ask if a central decision-maker should make allocations that are optimized for fairness within every round, or if fairness should be optimized across rounds. Freeman et al. (2017) analyzed (non-empirically) fair social choice in dynamic settings with many agents with changing preferences over multiple goods. They chose to optimize for the overall product of all agents' utilities, and investigated strategies from there. However, the question of whether real participants would optimize for global utilities across choices is undetermined, including whether they would optimize for additive utilities (the global *maxsum* solution) or the product of utilities (more similar to *inequality aversion*). Given previous research, we hypothesized that participants would behave differently when presented with the same choice repeatedly: in other words, we expected that the choices from Experiment 1 would not be repeated three times. We were interested in determining whether participants' choices would be described by the same metric in each round, or across all rounds, and whether these choices would change depending on the number of anticipated rounds. In summary, in Experiment 2 we presented participants with two or three repetitions of the same payoff matrices to determine how they would act, as central decision-makers, in the common real-world situation of providing a resource multiple times to the same group of people.

2.3.1 Method

Participants

Participants with U.S. IP addresses were recruited from Amazon Mechanical Turk for two additional conditions: "Repeated 2x" (n = 24, 0 participants excluded), and "Repeated 3x" (n = 21, 4 participants excluded). Participants were paid between \$2.50-\$3.00 for their participation. Participants were excluded if they failed the included attention check or indicated that they did not understand the experiment.

Stimuli and Procedure

In Experiment 2, the procedure and stimuli were similar to Experiment 1, except instead of viewing each matrix once, participants saw each matrix twice (in the "Repeated 2x" condition) or three times (in the "Repeated 3x" condition). Matrices were repeated an even ("Repeated 2x" condition) and odd ("Repeated 3x" condition) number of times to probe how participants would balance their choices. Procedurally, participants read the prompt from the "Nominal" condition, made a choice for the presented matrix, and justified their answer, as in Experiment 1. Then they saw the following prompt: "Your guests have finished the drinks, and you can now put out another. Which drink would you like to serve?" and were shown the same matrix again, and asked to make a choice and justify their answer. The latter prompt and matrix appeared once more in the "Repeated 3x" condition. Participants were able to scroll up and down the page of prompts and responses to determine the total number of drinks they were serving, and could change their choices at any time within the round.

Analysis

In Experiment 2, participants had the opportunity to serve multiple drinks. In the "Repeated 2x" condition they chose an option o^k from matrix $\mathcal{M}_m$ two times (o_1^k, o_2^k), while in the "Repeated 3x" condition they chose an option o^k matrix $\mathcal{M}_m$ three times (o_1^k, o_2^k, o_3^k).

Unlike in Experiment 1, choices from each repeated matrix were not independent given the matrix $\mathcal{M}_m$. We thus calculated two variants on individuals' preferred metrics. For the naïve analysis, we analyzed each choice in the repeated case as an independent decision. Specifically, we assigned all choices after o_1^k to new "independent" participants. Thus, in the "Repeated 2x" condition, the number of participants artificially doubled, with the first set of participants making choices o_1^k for each matrix, and the second set of participants making choices o_2^k for each matrix. $\mathcal{F}_q^{indep}$ was calculated and predictions were made using the same procedure as in Experiment 1, where *"indep"* indicates the artificial creation of independent participants.

For the more sophisticated analysis, we assumed that participants were making one large decision across two or three unordered matrices, rather than making semi-independent decisions o_1^k, o_2^k, (o_3^k). We hypothesized that $\mathcal{U}(o_1^k + o_2^k + o_3^k)$ would not be equivalent to $\mathcal{U}(o_1^k) + \mathcal{U}(o_2^k) + \mathcal{U}(o_3^k)$ (the assumption that choices were independent). We thus mapped the repeated choices to a more sophisticated non-repeated choice: one between all 10 (for the "Repeated 2x" condition) or 20 (for the "Repeated 3x" condition) combinations of individual choices o_1^k, o_2^k, (o_3^k) participants could have made. We then calculated $\mathcal{F}_q^{summed}$ based on the summed agent utilities across all of these potential

combination of matrices. Order of choices was not taken into account, but repetition of the same choice o^k for o_1^k, o_2^k, (o_3^k) was allowed. In short, by following this summing procedure, in the "Repeated 2x" condition, participants had 10 hypothetical choices, rather than the 4 they actually faced (two times). In the "Repeated 3x" condition, participants had 20 hypothetical choices, rather than the 4 they actually faced (three times). The analysis proceeded in the same way as for the "Nominal" condition in Experiment 1 except that the the matrices $\mathcal{M}_m$ expanded to contain 10 or 20 choices.

2.3.2 Results and Discussion

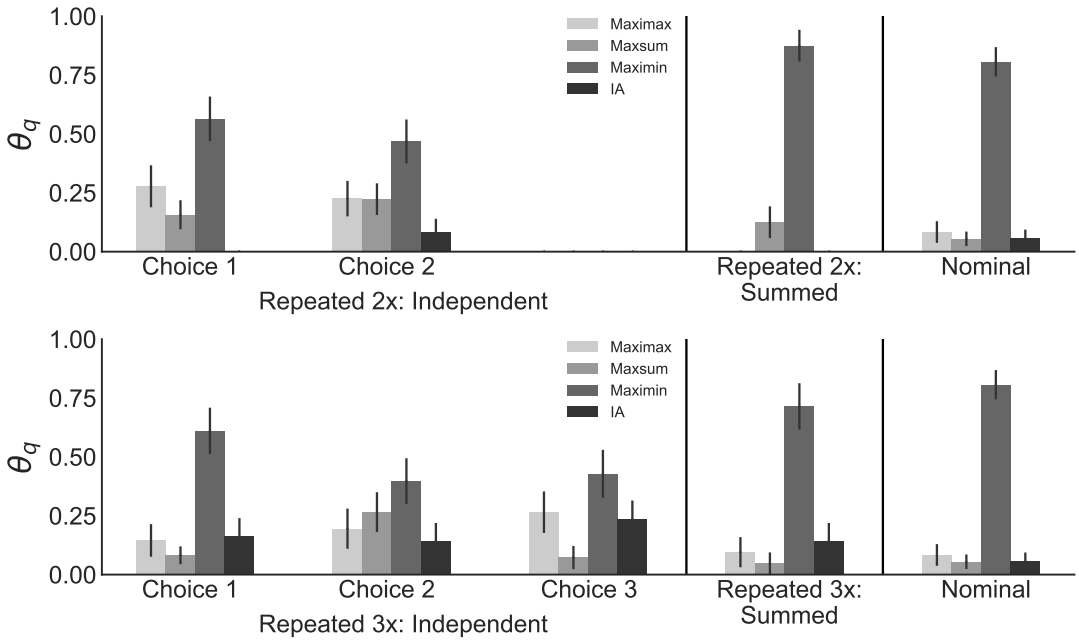


Figure 2.4: MaxEnt model results for Experiment 2, showing mean inferred weight $\theta \pm$ standard error for each metric across participants. **(Above)** Participants in the "Repeated 2x" condition, and "Nominal" condition. **(Below)** Participants in the "Repeated 3x" condition, and "Nominal" condition. From left to right: θ_q for each choice in the repeated condition (choices were treated as independent); θ_q for summed choices in the repeated condition, where new matrices were calculated according to $\mathcal{F}_q(o_1^k + o_2^k + ...)$; and θ_q from the "Nominal" condition.

In Experiment 2, we wanted to test how participant behavior would change across repeated conditions. One hypothesis was that participants would, as in Experiment 1, use the *maximin* metric for each individual choice, not keeping in mind the overall structure of the repeated choices. A second hypothesis was that participants would maintain *maximin* behavior when considering their

combined choices, but would engage in different behavior (for example, suboptimal behavior with respect to the *maximin* metric) within each individual matrix. A third hypothesis was that participants could have chosen to maximize the utility for one agent, then the other, alternating the optimal *maximax* options for each individual choice. This third hypothesis was not borne out in the data and so will not be further discussed.

Results especially supported the second hypothesis, that participants may not have acted independently according to the *maximin* metric on each choice, but rather made sure that the overall outcome after all actions was a *maximin*-supporting outcome.[2] To test this, we evaluated the metrics' values over participants' combined choices ($\mathcal{F}_q^{summed}$). Statistically, we compared participants' empirical $\mathcal{F}_q^{summed}$ values to 10000 samples of $\mathcal{F}_q^{summed}$ values from randomly-generated choices, where the values were summed over participants and matrices. This comparison yielded z-scores and p-values relating the empirical results to the null distribution for each metric. (The statistics here are the same as explained in Experiment 1). For both the "Repeated 2x: Summed" and the "Repeated 3x: Summed" conditions, the z-scores for the *maximin* metric were higher than for any other metric (Table 2.1), and were in the same general range as those of the "Nominal" condition. The statistical analysis thus supported the conclusion that participants' behavior could be explained by applying the *maximin* metric across the set of repeated matrices they were reasoning over. We also analyzed the results through the MaxEnt model. We found that the degree to which the *maximin* metric explained behavior in the "Repeated: Summed" conditions, both for "Repeated 2x: Summed" (two repeated choices) and "Repeated 3x: Summed" (three repeated choices), was high and almost equivalent to in the "Nominal" condition (Fig. 2.4).

Results also supported the first hypothesis to a lesser extent: that participants' behavior could be described by the *maximin* metric within each individual choice they made throughout the repeated choices. Here, we analyzed the data in terms of individual choices, completing a statistical analysis on the summed empirical and null distributions for the $\mathcal{F}_q^{indep}$ values. Z-scores were highest for the *maximin* metric for almost all conditions ("Repeated 2x: Independent (Choice 2)," "Repeated 3x: Independent (Choice 1)," "Repeated 3x: Independent (Choice 2)," and "Repeated 3x:

[2] As an example, if a participant in the "Repeated 2x condition were analyzing the matrix in Fig. 2.1, to maximize the *maximin* metric on each choice they would choose the [A: 5, B: 8] cup twice, but if they were maximizing the *maximin* metric across all choices they would choose the [A: 4, B: 11] cup once and the [A: 12, B: 2] cup once. Note that choosing the [A: 5, B: 8] cup twice is tied for the second-best option for maximizing *maximin* overall, so this set of choices would still contribute to the hypothesis that participants were maximizing the *maximin* metric across all choices, just not as strongly as would be true if the participant had chosen the [A: 4, B: 11] cup once and the [A: 12, B: 2] cup once.

Independent (Choice 3)") (Table 2.1). The exception was that the z-score for the "Repeated 2x: Independent (Choice 1)" condition was highest for the metric *maxsum*. To further analyze our data, we used the MaxEnt model to infer the combination of metrics that best described participant behavior over individual choices. We found that the *maximin* metric best described participant behavior for all of the conditions assuming independence: "Repeated 2x: Independent (Choice 1)," "Repeated 2x: Independent (Choice 2)," "Repeated 3x: Independent (Choice 1)," "Repeated 3x: Independent (Choice 2)," and "Repeated 3x: Independent (Choice 3)" (Fig. 2.4). The first hypothesis that participants would act according to the *maximin* metric for each of their individual choices was thus supported, though not uniformly across conditions, and contribution from a combination of metrics was visible within the MaxEnt results.

Supporting the importance of the *maximin* metric in describing behavior, estimating a single θ across all participants resulted in values dominated by the *maximin* metric for all but one condition in Experiment 2. Specifically, for the "Repeated 2x: Independent (Choice 1)" condition, $\theta_{maxsum} = 1$ and $\theta_{maximin}, \theta_{maximax}, \theta_{IA} = 0$, the sole condition where the *maxsum* metric was ranked highest. For the "Repeated 2x: Independent (Choice 2)" condition, $\theta_{maximin} = 1$ and $\theta_{maximax}, \theta_{maxsum}, \theta_{IA} = 0$. For the "Repeated 2x: Summed" condition, these averages were: $\theta_{maximin} = .81$, $\theta_{maxsum} = .19$, and $\theta_{maximax}, \theta_{IA} = 0$. For the "Repeated 3x: Independent (Choice 1)" condition, these averages were: $\theta_{maximin} = .92$, $\theta_{maxsum} = .08$, and $\theta_{maximax}, \theta_{IA} = 0$. For the "Repeated 3x: Independent (Choice 2)," "Repeated 3x: Independent (Choice 3)" and the "Repeated 3x: Summed" conditions, these averages were $\theta_{maximin} = 1$ and $\theta_{maximax}, \theta_{maxsum}, \theta_{IA} = 0$.

This pattern of results suggests that participants consider all of their choices and maintain a running *maximin* metric calculation, though they also seem to apply the *maximin* metric to a lesser degree within each individual choice. Support for this argument draws from the following observation: the dominance of the *maximin* metric in explaining behavior is greater for conditions in which choices were considered cumulatively (the "Repeated: Summed" conditions), than for conditions in which each choice was considered independently (the "Repeated: Independent" conditions). Specifically, the *maximin* z-scores were higher in the statistical analyses and the *maximin* θ values were higher in the MaxEnt model for the "Repeated: Summed" conditions compared to the "Repeated: Independent" conditions, and were in fact very similar to those in the "Nominal" condition.

Finally, to verify that our weight vectors in the MaxEnt model generalized, we used participants' θ vectors to predict held-out data. Training and testing on 50% participants and matrices, the "Repeated 2x: Independent" condition had 47.0% predictive accuracy (non-held out comparison: 53.6%, chance accuracy: 25%); the "Repeated 2x: Summed" condition had 39.3% predictive accu-

racy (non-held out comparison: 43.3%, chance accuracy: 10%); the "Repeated 3x: Independent" condition had 45.8% predictive accuracy (non-held out comparison: 57.9%, chance accuracy: 25%); the "Repeated 3x: Summed" condition had 10.9% predictive accuracy (non-held out comparison: 17.6%, chance accuracy: 5%). Our weight vectors were less accurate in predicting "Repeated: Independent" conditions than they were compared to the non-repeated conditions, but had relatively high predictive accuracy for held-out test sets compared to non-held-out test sets. In the "Repeated: Summed" conditions, as chance accuracy fell, predictive accuracy also fell. With such low predictive accuracy in the "Repeated 3x: Summed" condition especially, we sought to test whether this was an artifact of the many potential choices available to participants in the "Repeated" conditions or a consequence of the utilities of the choices available, motivating Experiment 3.

We observed that in both the "Repeated" conditions and the conditions from Experiment 1, participants' behavior was best described by the *maximin* metric. The lack of difference in the Repeated conditions could be construed as a null result, under the assumption that we did not manipulate the multi-decision conditions strongly enough to cause a change in participants' responses. To show that participants were producing different behavior in the "Repeated" conditions, but nevertheless overall their behavior could be best described by the *maximin* metric, we show that there is a different pattern of responses for each "Repeated" independent choice condition compared to the "Nominal" condition. Specifically, we computed the percentage of matrices wherein participants had significantly different responses between conditions. We found that that percentage was lower in comparing the "Repeated (Choice 1)" and "Nominal" conditions, and higher in comparing the "Repeated (Choice 2)" and "Nominal" conditions (Table 2.3). This percentage also changed with the "Repeated 3x: (Choice 3)" and "Nominal" condition comparison (Table 2.3). These percentage differences indicate that participants were making different choices within each matrix of the "Repeated" conditions, even while the overall results from the MaxEnt model showed continued *maximin* behavior.

Conditions	*# sig.*	*# inc. total*	*% sig.*
Rep2x: Ind.(C1) vs. Nominal	7	20	35%
Rep2x: Ind.(C2) vs. Nominal	12	20	60%
Rep3x: Ind.(C1) vs. Nominal	0	17	0%
Rep3x: Ind.(C2) vs. Nominal	15	20	75%
Rep3x: Ind.(C3) vs. Nominal	9	19	47%
Rep2x: Ind.(C1) vs. (C2)	4	19	21%
Rep3x: Ind.(C1) vs. (C2) vs. (C3)	7	20	35%

Table 2.3: Results showing differences in participants' choices across conditions. Shown is the percentage of matrices ("% sig.") wherein a χ^2 test of independence showed the histograms of participants' choices for each matrix were significantly different ($p = .05$: uncorrected) across the listed conditions. In more detail: for each matrix, the number of participants who picked each choice was summed, and this set of values was compared across the two experimental conditions listed. If a matrix had any expected value of 0 in the computed χ^2 frequencies, that matrix was removed from the analysis. Note that expected values were often < 5. The number of matrices that were significantly different according to the χ^2 test of independence ("# sig.") was divided by the total number of included matrices ("# inc. total"). Note that condition names are shortened: "Rep2x: Ind.(C1)" represents the "Repeated 2x: Independent (Choice 1)" condition.

2.4 Experiment 3: Summed repeated choices

When taking repeated actions, participants' individual choices tended toward being described by the *maximin* metric, but were described by a combination of other metrics as well. Why were participants' repeated choices not as clearly described by the *maximin* metric as they were when making a single choice? It could be that participants were attempting to view all of the repeated trial as a single decision, and were trying to maximize the *maximin* solution across all choices, but that the mental overhead for this computation led them to less clearly *maximin* solutions. Alternatively, perhaps computational overhead was not very influential in participants making different choices than they had in Experiment 1, and there was something about the way that the repeated stimuli were presented that led to dissimilar results in Experiment 1 and 2—for example, perhaps participants felt pressure to vary their choices in Experiment 2 when faced with the same questions.

To isolate whether computational overhead was an effect, we presented participants with the summed versions of the repeated choices they saw in Experiment 2. Participants could then see the summed version of choices in Experiment 2, so the computation was easier if they were attempting to sum utilities across decisions. This manipulation also allowed us to present participants with a slightly different one-shot matrix (with six options rather than four) to see if the results from Experiment 1 generalized. We investigated whether participant choices would be similar to those in Experiment 1, individual decisions in Experiment 2, and the artificially-summed decisions in Experiment 2, with an emphasis on whether participants' behavior would be more or less clearly described by the *maximin* metric, as we expected from Experiments 1 and 2. In summary, we simplified the multi-decision problem presented before, testing what metrics described participants' solutions when Experiment 2's repeated choices were condensed into a single decision again.

2.4.1 Method

Participants

Participants with U.S. IP addresses were recruited from Amazon Mechanical Turk across two additional conditions: "Follow-Up (2x)" (n = 31, 0 participants excluded), and "Follow-Up (3x)" (n = 29, 3 participants excluded). Participants were paid between $2.50-$3.00 for their participation. Participants were excluded if they failed the included attention check or indicated that they did not understand the experiment.

Stimuli and Procedure

The stimuli and procedure were similar to Experiment 1. Participants again viewed each matrix once, but each matrix contained 6 choices that were drawn from the "summed" utilities from Experiment 2. In Experiment 2, for the "Repeated 2x" experiment, participants had 10 hypothetical choices per round, and in the "Repeated 3x" experiment, participants had 20 hypothetical choices per round. These hypothetical choices (of the form [A's utility, B's utility]) were what we used to generate each matrix in Experiment 3.

To generate each matrix, we first tried to include one choice that maximized each metric. Specifically, we examined all 10 or 20 choices for each round from Experiment 2, and for each metric selected the choice that would be most preferred under that metric (e.g. [4, 4] might be chosen under the inequality aversion metric, but not the maxsum metric if [8, 4] was also an option in the set). If two metrics had the same best choice (irrespective of order, so [8, 4] was identical to [4, 8]), this choice was only included once. If a metric had several best choices (e.g. under the inequality aversion metric, [3, 3] would be as good as [4, 4]), then an unused choice was selected at random. (Metrics with one best choice were examined first, then metrics with multiple choices were examined in random order.) The remaining choices, since 6 choices were included in each matrix, were selected from the most popular unused choices from Experiment 2.

2.4.2 Results and Discussion

In Experiment 2, we observed that participants appeared to be reasoning over the whole set of repeated choices, and that their behavior was best described by the *maximin* metric across that set ("Repeated: Summed" conditions). To further test this hypothesis, we constructed matrices where participants only had to make one choice, but each choice constituted the sum of previously repeated choices in Experiment 2. Our results corroborated those from Experiments 1 and 2. With these single-choice matrices, participants' behavior could be best explained by the *maximin* metric. In a statistical analysis comparing the empirical choices made by participants to a simulated set of random choices, z-scores were highest for the *maximin* metric as compared to the other metrics in the "Follow-Up (2x)" and "Follow-up (3x)" conditions (Table 2.1). In our MaxEnt model, participant behavior was also best explained by the *maximin* metric in the "Follow-Up (2x)" and "Follow-up (3x)" conditions, to a degree similar as in the "Repeated: Summed (2x)" and "Repeated: Summed (3x)" as well as the "Nominal" conditions (Fig. 2.5).[3] The similarity of values across the

[3] Our results from estimating a single θ across all participants, $\theta_{maximin} = 1$ and $\theta_{maximax}, \theta_{maxsum}, \theta_{IA} = 0$ for the "Follow-Up (2x)" and "Follow-Up (3x)" conditions, also support

"Follow-Up" conditions and the "Repeated: Summed" conditions supports the hypothesis described in Experiment 2, that participants consider the whole set of choices when they make decisions rather than just each choice individually.

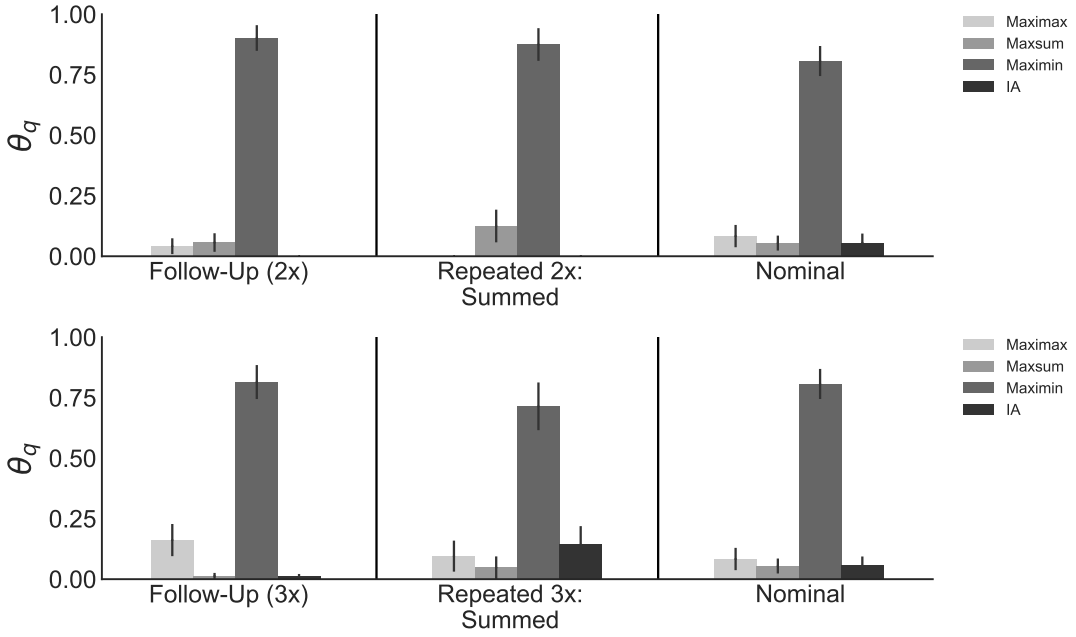


Figure 2.5: MaxEnt model results for Experiment 3, showing mean weight $\theta \pm$ standard error for each metric across participants. **(Above)** Participants in the "Follow-Up (2x)," "Repeated 2x" (summed choices, where new matrices were calculated according to $\mathcal{F}_q(o_1^k + o_2^k + ...)$), and "Nominal" conditions. **(Below)** Participants in the "Follow-Up (3x)," "Repeated 3x" (summed choices), and "Nominal" conditions. The *maximin* metric best described participant behavior.

Finally, to check that our MaxEnt weight vectors generalized, we used participants' θ vectors to predict held-out data. Training and testing on 50% participants and matrices, the "Follow-Up (2x)" condition had 71.6% predictive accuracy (non-held out comparison: 71.6%, chance accuracy: 16.7%); the "Follow-Up (3x)" condition had 60.5% predictive accuracy (non-held out comparison: 66.7%, chance accuracy: 16.7%). We could predict participant responses to the "Follow-Up" conditions well with and without held-out testing sets, indicating the strength of our inferred weight vectors. The high predictive accuracy in the "Follow-Up" conditions compared to the "Repeated: Summed" conditions indicates that the previous low predictive accuracy was a consequence of the large number of possible choices available in the "Repeated" conditions, and the complexity of not

the conclusion that participants' behavior was best explained by the *maximin* metric.

seeing the end results directly and having to compute them, rather than a consequence of the utilities of the choices themselves.

2.5 General Discussion

If a schoolteacher is trying to choose a field trip for a group of students with kinetic learners, visual learners, children who don't speak English at home, boisterous students, and students who are scared of new places, what should they do? How should people donate their money, governments choose between aid programs, or assistive robots mediate between family members' preferences? In this work we examined the broader question of how a decision-maker should act when people have different preferences. Specifically, we asked people what *they* would do in a paradigm where they could take only one action, making a single drink, to the inevitable dismay of some of the people they were serving. While it is not clear that decision-makers should automatically match the behaviors that people intuitively use to make group-level decisions, it is informative to know the ground truth of what people feel are good decisions.

We analyzed participants' choices by assuming their behavior could be captured by a combination of four metrics. Of these metrics—*maximax*, *maxsum*, *maximin*, and *inequality aversion (IA)*—we observed that participants' behavior could be reliably described by the *maximin* metric, the idea of maximizing the utility of the worst-off agent. With respect to our story, the schoolteacher may not know what's the right thing to do, but we at least know how people behave: for each field trip find the child who would have the worst time, and choose the field trip wherein that child enjoys the field trip as much as possible.

Participants behaved according to the *maximin* metric despite changes in the agents involved. Specifically, whether considering a robot manager, human manager, or serving friends or strangers, participants' behavior was similar and described by the *maximin* metric (Experiment 1). Participants did tend to move towards behavior more described by the *maxsum* metric when they were told they were beneficiaries of the decision, perhaps reasoning about the task more as a gambling situation and less as one mandating fairness for unknown recipients. In Experiment 2, we found that when participants made repeated decisions, each individual choice therein was characterized by a modest *maximin* preference. However, participants also acted as if they maintained a running total of their choices across repeated decisions. In particular, participants' behavior was best described by the *maximin* metric under the assumption that they were summing utilities across repeated decisions. Experiment 3 provided additional support for the hypothesis that people maintain an overall calculation in repeated decision-making, rather than reasoning over each problem individually.

When participants' cumulative choices from Experiment 2 were summed to create a combined set of matrices for Experiment 3, participants' behavior was described by the *maximin* metric to the same degree as was shown in the "Repeated: Summed" and "Nominal" conditions. All of these results suggest that participants keep track of calculations over time and prefer making decisions consistent with the *maximin* metric when allocating indivisible items across agents.

In the remainder of the paper, we discuss the relationship of these results to previous work, and ways in which these results could be extended.

2.5.1 Relationship to previous work

We asked what a decision-maker should do when it could take only one action—create one resource—to be shared among multiple other agents, and our results support the hypothesis that participants prefer to behave in ways described by the *maximin* metric. How, then, do our results compare to the literature? In the nearby literature of fair allocation, there is not nearly so neat a consensus. However, in most fair allocation studies the participant acts as a self-interested party, weighing the desires of selfishness and "fair" allocation. We are interested in the case of what people consider helpful actions when they have no stake in the outcome. Four studies are similar to our study in that respect.

The first is Herreiner and Puppe (2007), who presented participants with payoff matrices in which agents could receive different "goods." Unlike our study, in which there was only one item to be created / allocated, the paradigm in Herreiner and Puppe (2007) resembled an estate game in that multiple items, the number of which was often larger than the number of agents, could be distributed to various agents. This added complication to the proceedings, as participants then had large numbers of item combinations to reason over (e.g. their Problem 1 had $3^3 = 27$ different allocations), and they could consider fairness not only over utilities, but over "bundles"—the set (and number) of items allocated. To this end, Herreiner and Puppe (2007) constructed their payoff matrices with an eye to the metric of "envy-freeness," which describes whether any given agent would be *envious* of any other agent's bundle.

The results from Herreiner and Puppe (2007) seem to be consistent with a hypothesis that participants prefer acting according to a *maximin* metric, given that in their study this metric was sometimes confounded with different metrics within a single choice. However, in their Problem 1 and Problem 6, Herreiner and Puppe (2007) showed results indicating that participants preferred the *IA* metric over the *maximin* metric. In Problem 1, participants chose to withhold the last item rather than give one agent two items compared to the other's one (preferring the two agents to have

utilities [49,48] rather than [49,53]), and in Problem 6, participants distributed four items across three agents to maintain exactly equal utilities [45,45,45] rather than choosing the best *maximin* allocation [48,60,52]. Problem 1 is an interesting case in that participants were considering fairness in both utilities *and* item number. Our participants tended to choose according to the *maximin* metric when only considering utilities, but it could be that if we had asked them to reason over both utilities and item number they would have considered choices emphasizing inequality aversion as fairer and better, especially when the difference in utilities was small. In Problem 6, our results suggest that choice complexity could have been influencing the observed results in Herreiner and Puppe (2007), and in future work it would be interesting to check if participants would choose the *IA* solution [45,45,45] rather than the *maximin* solution [48,60,52] if directly presented with those choices, rather than being asked to add utilities from four items across three agents. As a final note on the complexity of addition, Herreiner and Puppe (2007) noted in their discussion that participants' final solutions were correlated with their reported allocation procedures—e.g., the order in which participants assigned items to each agent. Allocation procedures were not inherent to our problem setup but are well-considered within the fair allocation literature (see e.g. Dupuis-Roy and Gosselin (2011)), and perhaps also add implicit utilities to participants' preferred choices.

In Engelmann and Strobel (2004), the basic structure of the task was very similar to ours, as participants made choices between three items that three agents had utilities over. Though the participant acted as one of these agents, the participant's utility was always held fixed over all items. Engelmann and Strobel (2004) focused their payoff matrices on distinguishing between two existing models, each of which had one utility function designated as participants' preferred allocation metric. As such, Engelmann and Strobel (2004) also had many matrices that confounded the metrics we considered, but with this caveat the *maximin* metric could be considered a primary motivation for participants' behavior. Relevantly, Engelmann and Strobel (2004) state that one of the models they evaluated performed well because it captured the *maximin* metric, but that overall a combination of *maximin*, *maxsum* (which they call efficiency), and *selfishness* (which we do not consider) considerations drove participant behavior. Indeed, in several of their payoff matrices, both the best *maxsum* and *maximin* choices were often selected by participants, and the *maxsum* choice often garnered a higher proportion of participants.

E. Fehr et al. (2006), however, provides a rebuttal to the Engelmann and Strobel (2004) paper, replicating the Engelmann and Strobel (2004) study with non-economics participants and showing that the *maxsum* solution was selected far less by participants who were in different programs. E. Fehr et al. (2006) did not choose payoff matrices that distinguished between the *IA* and *maximin* metrics, but their results hint that the subject pool can influence preferred allocation metrics.

Herreiner and Puppe (2007) tested economics and law student participants, but our study was conducted on Amazon Mechanical Turk and had a broader population than economics undergraduates. In future work, it would be interesting to replicate our study with economics undergraduates, and observe whether this difference is enough to observe participants' preference for the *maxsum* metric over the *maximin* metric. Fairness perceptions have been observed to differ across different populations (see, e.g. Andreoni and Vesterlund (2001), Croson and Gneezy (2009), Gaertner and Schwettmann (2007), Marwell and Ames (1981), and Camerer (2011) summarizes some demographic results for behavioral game theory) so this hypothesis would not be improbable, though the review in Konow and Schwettmann (2016) cites that generally demographic variables seem to have relatively small effects on economics experiments.

Yaari and Bar-Hillel (1984) presented several different types of scenarios wherein a third party allocated goods according to the *maximin* metric, *maxsum* metric, and *inequality aversion* metric. Participants played three types of games, where they had to allocate fruits according to receiving agents' *needs* or *tastes / utilities*, and also across agents' differences beliefs about the fruits. Participant behavior differed across the varying conditions, shifting according to the tradeoffs presented, as in this work. The authors observed that when choices were presented in terms of needs—one agent needs a certain type of vitamin to be healthy, so they need a certain type of fruit—82% of participants chose the *maximin* allocation. In our paradigm, agents did not have needs, rather stated utilities (preferences over choices). This was much closer to the second condition in Yaari and Bar-Hillel (1984), in which agents had different "tastes" or stated utilities. Intriguingly, participants did not adhere as strongly to the *maximin* choice in this case, as instead 28% chose the *maximin* solution, and 35% of participants chose the *maxsum* allocation. The distribution of choices also changed in the third condition of Yaari and Bar-Hillel (1984), in which agents had different perceptions of how much value each of the items would give them. Yaari and Bar-Hillel (1984) concluded from their experiments that the *maximin* metric best describes participant behavior, but only when needs are salient.

In our work, we found that participants made choices according to the *maximin* metric outside of needs-based formulations and did so consistently across changes in wording and repetition over time. Our paradigm differed from that of Yaari and Bar-Hillel (1984), however, which could account for the difference in results. The main difference between our work and that of Yaari and Bar-Hillel (1984) is that we asked a different question: given agents had utilities over a set of four or six options, which option should be chosen to best satisfy those utilities? Yaari and Bar-Hillel (1984) asked a fair allocation question: given a specific number of items (and agents' preferences over the different items in the "tastes" condition), how should those items be divided between agents? The

fair allocation paradigm is necessarily zero-sum, wherein if one agent receives a resource, another loses it. In particular, in question 4 from Yaari and Bar-Hillel (1984), participants were asked to divide 12 grapefruit and 12 avocados between two agents. One agent was stated as hating avocados (utility = 0), but would buy grapefruit if they were priced under $1 per pound. Meanwhile, the other agent liked both avocados and grapefruit and would pay for them if they were priced under $.50 per pound (half the utility of the first agent for grapefruit). With these agent preferences in mind, participants were now expected to divide a fixed number of grapefruit and avocados. Compare this to our study, in which participants decided which drink to give to two agents to share. This task is not zero-sum—if one agent has high utility for a drink, this does not detract from another agent's utility for that drink. We did, however, have four matrices out of 20 for each experiment in which the joint utilities for each drink were identical, meaning that no matter what drink the participant chose, the agents would only jointly achieve a given happiness. But in this scenario the question was not "how many of each item should be given to each agent given their utilities" (as in Yaari and Bar-Hillel (1984)), but "given there is a fixed amount of utility to be had, how should that utility be divided between agents?" This is a distinct question, and it is probable that participants reason differently about a zero-sum problem of "distribute 12 pieces of fruit between two people or throw some away" (allocation) compared to "create one shared item that will make two people more or less happy." We may expect that Yaari and Bar-Hillel (1984) did not observe as much *maximin* behavior because they used a task of zero-sum bundle allocation; this hypothesis should be investigated in future work.

Another difference between our study and that of Yaari and Bar-Hillel (1984) was that we probed participants' intuitions of what was fair with 20 questions (4 or 6 choices each) for each experiment, and determined which metrics best described participants' behavior by accumulating evidence across all of these trials. Alternatively, Yaari and Bar-Hillel (1984) had participants answer one question for each experiment (usually with 5 choices), and used that single choice to inform which single "mechanism" (analogous to our metrics, but without considering combinations of metrics) best described participant behavior. The authors thus did not have a continuous characterization of how much participants were acting in accordance with different metrics, instead using discrete choices that distinguished "*maximin*," "*maxsum* / utilitarian," and other allocations, which they tested with a single set of choices. With more trials available and a finer-grained measure of how each choice contributed to the various metrics, Yaari and Bar-Hillel (1984) may have observed a more *maximin*-biased distribution, as we did.

As a final aside, with regards to changing in wording, Yaari and Bar-Hillel (1984) found similar distributions of responses when they asked, first, how participants would divide the items, and

second, how the two agents would divide the items if the agents were aiming to be just. In our paradigm, we asked only what a third-party decision-maker would do, but this result implies that in our paradigm we would find similar responses if we were asking participants what choices recipients would think were just.

Our study stands in contrast to these four studies in a few ways. First, while our problem formulation can encompass any fairness allocation problem based on our definition of an "action," the specific paradigm we used was specialized to solve the problem of what single item a decision-maker should choose to distribute to a group (a rather straightforward action). This stands in contrast to fair allocation studies, and we described a direct comparison of our study and Yaari and Bar-Hillel (1984) on this axis, with a particular eye to zero-sum choices. Additionally, some differences in our results may be attributable to experimental simplicity. For example, in Herreiner and Puppe (2007), participants may entertain additional implicit utilities when they reason over bundles (such as item number fairness), and there may be difficulties in balancing many possible choices while considering different allocations of 3-4 goods. Similarly, Yaari and Bar-Hillel (1984) introduce motivational features not present in our work, like division based on agent needs. A second methodological distinction of our study is that the other studies did not use a continuous measure of behavioral alignment to several metrics, instead using discrete comparisons over metrics that largely correlate, which leaves the option open that finer-grained distinctions may change their details of some of their results. Additionally, a significant advantage of our study is that we employed 20 matrices and could have employed more, whereas other studies had fewer than 12 payoff matrices, and these payoff matrices often confounded different metrics due to the authors' different emphases. Finally, we showed consistent participant behavior aligning with the *maximin* metric, rather than emphasizing the involvement of the *maxsum* and *IA* metrics, and observed this finding over repeated conditions, a contrast which has not been previously conducted to the authors' knowledge.

In summary, our study is aimed at a different question than most in the fairness literature; we aim to study how people would prefer that decision-makers act when they can benefit multiple people. This question is not addressed in previous studies, but is most related to other fairness studies examining uninterested third-party (zero-sum) decisions (note that unlike a fair allocation problem, our problem is not zero-sum since a decision-maker is creating a resource for two people with different but not opposing utilities.) Within this previous work, preference for the *maximin* metric has not been universally shown, and the *maximin* metric is often not directly compared with other potential metrics. Relatedly, previous studies often do not account for the correlations among metrics when describing participant choices. Here, we presented many choices to participants, testing their intuitions many times for each question, and then used a MaxEnt model to disambiguate between

the overlapping metrics that described each choice. We additionally probed participants' behavior in repeated, longer-term scenarios than have been evaluated for this problem setting. We distinguished participants' choices representative of the *maximin* metric by direct comparison with competitive strategies, across many question formulations to ensure generalization. We thus provide a novel contribution to the question of what people think a decision-maker should do when faced with helping people with unique utilities.

2.5.2 Future directions

An interesting extension to this work would be to do the full study that includes extreme comparisons. If participants are faced with the choices [4, 80] and [4, 4] (where agent A receives the first number and agent B receives the second), all participants will likely choose [4, 80], because the tradeoff between maximizing the sum and trying to maintain equality between agents is so extreme. These choices could be varied parametrically, gradually making the tradeoffs less extreme (and more difficult to choose between) with respect to different metrics. Our study lies basically at the center of that parametric descent, where the choices are most difficult. It is difficult to choose whether [4, 8] or [5, 5] is better, and behavior according to both the *maximin* and *maxsum* metrics are competitive. In this range, determining whether a participant was acting more according to a *maximin* or an inequality aversion metric required accumulating evidence across trials, motivating the use of our MaxEnt model. This is because each choice that a participant made could serve a few possible metrics simultaneously, and in Table 2.1 we observed this by noting that in the raw data, participants' behavior tended to be described by the *maxsum*, *maximin*, and inequality aversion metrics rather than a single metric. However, our paradigm would work well in evaluating when participants would change their behavior as tradeoffs become more extreme. If a participant acts according to the inequality aversion metric until the group utility hits a threshold, that participant should then continue acting according to a *maxsum* metric for all the more extreme choices thereafter.

Tradeoffs like those between equity, the *maxsum* metric, and need (which we don't examine here), have been widely compared in previous studies. Work like Ahlert et al. (2013), Charness and Rabin (2002), Engelmann and Strobel (2004), Faravelli (2007), E. Fehr et al. (2006), Fisman et al. (2007), Konow (2001, 2003), Konow and Schwettmann (2016), Mitchell et al. (1993), Ordóñez and Mellers (1993), Pelligra and Stanca (2013), Schwettmann (2009, 2012), Skitka and Tetlock (1992) find that participants do trade off between different principles based on the choices presented, as we see in our work when participants make choices in accordance with several of the metrics. The suggested set of experiments would confirm and expand upon our results, as our stimuli were chosen to

isolate which metrics participants thought best when choices were most difficult. Moreover, a parametric study examining these tradeoffs would provide a large and systematic dataset to contribute to the empirical literature on this topic.

Another extension to this study could be to consider whether having verbal problem descriptions, or other representations of utilities, would enhance our paradigm. Hurley, Buckley, Cuff, Giacomini, and Cameron (2011) replicated the study by Yaari and Bar-Hillel (1984) with quantitative problem descriptions, verbal problem descriptions, and both combined. In their verbal descriptions, they instructed participants to create an allocation according to a given metric: "Divide the apples in such a way that the total amount of vitamin F obtained by both Jones and Smith together is as large as possible," is an example of a *maxsum* instruction. The authors report that the results from the verbal problem descriptions were consistent with participants not understanding the relationship between the described metrics and their quantitative allocations, and that participants increased *maxsum* behavior. We used quantitative information in the present study; it would be hard to adapt the flexibility inherent in the different choices we presented to participants in a verbal-only description. Hurley et al. (2011) find that verbal and quantitative descriptions together produce results that more closely match quantitative descriptions alone. Because verbal descriptions do not seem to produce an increased understanding, we expect our paradigm works well with quantitative information, though the question of how to represent utilities and the effects of that choice on participant behavior remains an interesting one.

One limitation of this study is that it asks participants to evaluate between two agents only. The fair allocation literature commonly evaluates over two or three recipient agents; we expect our results to also scale to three agents, but we do not know how robust our findings will be at larger group sizes. This question is incredibly important given that large-scale decision-making impacts many people and so should be done in a way endorsed by many. We do not know that our results will scale: as group size increases, utilities become harder for participants (and people in general) to reason over, so we expect different heuristics will emerge. Fortunately, our experimental paradigm would serve as a good platform for this work given that we can easily generate sets of matrices which require participants to make difficult decisions, and can analyze participants' responses with various insertable metrics.

Finally, in future work, we hope to expand the generalizability of our findings. Within our paradigm, we observed consistent and robust results which were supported by using many matrices and conditions. Our paradigm differed from most in the fair allocation literature due to its focus on a more general problem: what a decision-maker should do when it can take one action (in this case making a drink) that will be shared among agents with different preferences. This paradigm itself offers

many avenues for expansion—there are many other actions a decision-maker could take besides making an item, including choosing a field trip, donating to organizations, or making governmental policy decisions. We should also consider actions that produce choices with negative utilities, since the dynamics at play in e.g. trolley problems or risk or loss aversion will likely lead to different behaviors. Drawing from the fair allocation literature also shows that this paradigm could be made more complex as multi-step actions are introduced (e.g. an action encompassing bundle distribution), or selfishness, need, desert, and other factors are considered. Konow (2003) presents a pluralistic approach, in which the prime considerations when thinking about justice are a sense of equality and need, utilitarinism and welfare economics, equity and desert, and context (other papers that present pluralistic approaches are Cappelen, Hole, Sørensen, and Tungodden (2007), Deutsch (1985), Frohlich and Oppenheimer (1992), Konow and Schwettmann (2016), Lerner (1975), Leventhal (1976)) and all will need to be incorporated into a fuller model of what people consider desirable behavior.

This work provides a general problem framework and quantitative model of what people think a third party should do when taking an action that will bring people different utilities. While the problem we address is distinct from that of fair allocation, it is informed by this literature and can contribute to the discussion of desirable descriptions of helpful behavior. Understanding what actions to take when every choice affects the well-being of many people with disjoint preferences has important bearings on the future, whether in designing decision-making algorithms for artificial intelligence systems, quantitatively evaluating how to help individuals make decisions that are better in line with what people consider to be good choices, or even creating accepted assistive robots. This and future work focusing on making our results more generalizable—and determining whether we would like our decision-makers to adopt people's intuitions of preferred choices—will provide important contributions to this problem.

2.6 Conclusion

Decision-making would be a much easier problem if everyone had the same preferences. Even having everyone agree about what to do about diverging preferences would help reduce the complexity of reasoning over many diverse perspectives. While that's not quite the spirited environment we live in, in today's world we have an opportunity to make these complex decision-making tasks easier with artificial intelligence systems. Artificial intelligences can optimize any number of people's preferences once they know what they are, and in this work we develop a quantitative model of people's intuitive preferences for what to do when making difficult decisions to help people.

Determining what decision-makers should do is a topic that philosophers and governors have

wrestled with for centuries: how do we think about inequality within individuals and in society? In this work, we ventured into the morass, developing a general problem framework that let us ask people how they would choose one action whose impact would be different for everyone. Deeper questions remain: what are humanity's morals around inequality, and how does what we do compare to how we think we should be? We don't know what to teach our decision-makers yet, and determining the answers to these questions will require significant collaborative effort. Fortunately, we did determine what drinks to serve at the resulting conferences.

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 longer 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 Introduction

It's a quiet afternoon, and you and a new friend are trying to entertain yourselves. "Hiking? Movie?" you suggest. "Movie," she says immediately. Feeling imaginative, you offer another choice: "Movie or science museum?" She takes longer to think on this. "...movie," she says eventually. Taking into account how long it took your friend to make her choices, you might guess that she hates hiking but science museums are a good future activity.

When we observe others making choices, we learn from the concrete choices they make, but also from cues like body language and timing. Inferring preferences from response times has a long history in economics and psychology (e.g. Busemeyer, 1985; Busemeyer & Townsend, 1993; Chabris, Laibson, Morris, Schuldt, & Taubinsky, 2008; Chabris, Morris, Taubinsky, Laibson, & Schuldt, 2009; Diederich, 1997, 2003; Gill & Prowse, 2017; Moffatt, 2005; Spiliopoulos & Ortmann, 2018; Wilcox, 1993). Recent sequential sampling models solve this problem by thinking of response time as an observable measure of a noisy cognitive process, where evidence is accumulated over time and a choice is made between options when the integrated evidence crosses a threshold. The longer it takes to make a decision, the closer those items are in the decision-maker's preferences, and the harder it is for the decision-maker to differentiate them. This principle has driven the use of sequential sampling models, the most common of which is the Drift Diffusion Model (DDM), to infer decision-makers' preferences from their response times (Bogacz, Brown, Moehlis, Holmes, & Cohen, 2006; Busemeyer & Rieskamp, 2014; Ratcliff, 1978; Ratcliff & McKoon, 2008; Ratcliff, Smith, Brown, & McKoon, 2016). Previous work has established that models such as the DDM can be used to infer people's preferences from their response times (see section "Background and previous

work" below).

Here we ask a different question: Can people infer *other people's* preferences based on observing their response times and choices, modeling those others' decision processes as a diffusion process? In modeling terms, the DDM predicts a decision-maker's response times and choices given their preferences. Can we capture observers' inferences using a rational model in which we invert the DDM to predict decision-makers' preferences (Figure 3.1)?

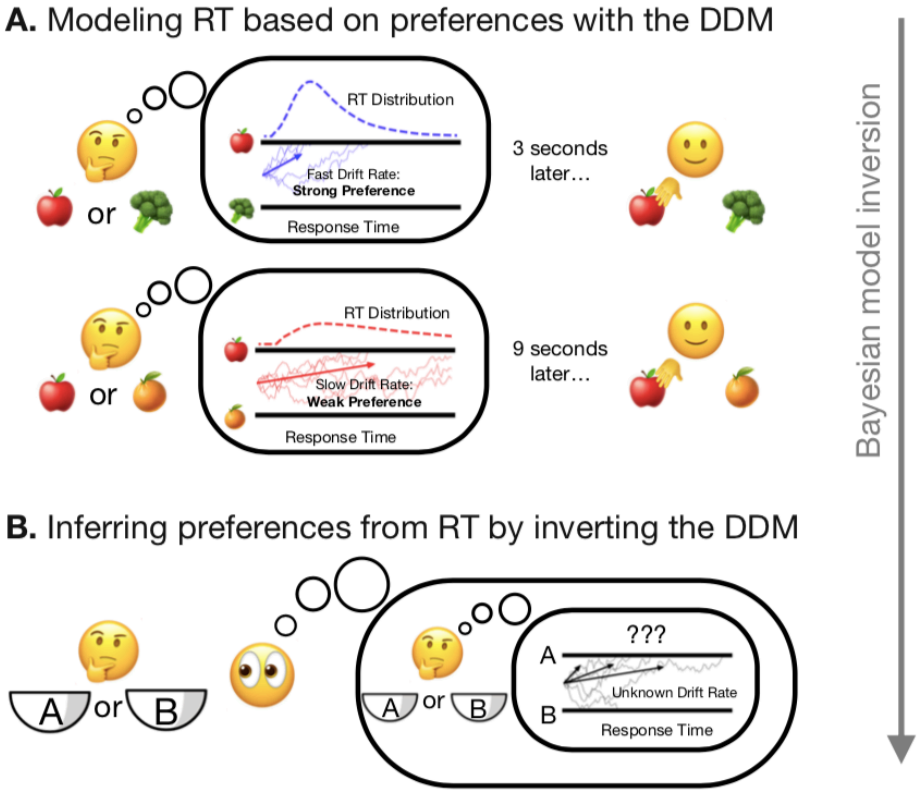


Figure 3.1: Schematic describing the Drift Diffusion Model (DDM) and the inverted DDM used in this paper. (A) The DDM is a generative model which predicts people's response times (RT) and choices given their preferences across options. Parameters include the drift rate $\mu = \beta(u_a - u_b)$ (value difference, or strength of preference; u_a and u_b are the utilities for the two items), threshold θ (carefulness), and the drift multiplier β. The stronger the preference for an item, the faster the decision. (B) In this work, we invert the DDM to generate predictions for the inference task: people inferring others' preferences based on observing their response times and choices.

People make inferences about preferences from response times constantly; using your friend's response times and choices to infer her preference for science museums over hiking is one illustrative example. People can make even more complex inferences from response times and choices if

they know the decision-maker's mental state, like how carefully they are making a decision. Your grandfather knows that if someone's making a choice in a hurry, then their response times are less informative. He likewise knows that if someone's very cautious, then a fast decision is particularly meaningful. In this paper, we propose that these forms of everyday inferences can be formalized and described by inverting a generative process like the DDM.

We explore the predictions of our inverse DDM through three experimental questions. In Experiment 1, we confirm the basic prediction that people will infer how much a decision-maker values one item over another based on observing the decision-maker's response times. In Experiment 2, we extend this analysis to testing people's predictions for a decision-maker's novel choice based on having previously inferred their relative strength of preferences. Just as you could infer a new friend likes science museums more than hiking based only on two indirect questions, our model suggests that participants will predict a decision-maker's choice in a novel comparison based on inferring the decision-maker's relative preferences from previous response times and choices. In Experiments 3, 4A, and 4B, we explore the possibility that participants will be able to infer even more from a decision-maker's response time if they are aware of the state of the decision-maker. When someone makes a quick decision, maybe it is because they value one choice much more than another, but maybe it is because they are feeling careless that day and are rushing through inessential decisions. These two factors correspond to different parameters of the DDM—"value difference" and "threshold"—which jointly predict the response time, and we ask if participants can perform joint inference over these parameters. We thus inform participants of the decision-maker's carefulness, and observe whether participants' predictions about their preferences vary.

3.2 Background and previous work

DDMs are often applied to perceptual decision-making tasks (e.g. Bitzer, Park, Blankenburg, & Kiebel, 2014; Brunton, Botvinick, & Brody, 2013). However, DDMs have also been used to model value-based decisions integrating both choice and timing information (Shadlen and Shohamy (2016) and Polanìa, Krajbich, Grueschow, and Ruff (2014) discuss relationships between DDMs describing perceptual versus value decision tasks). An important property of DDMs applied to value-based decisions is that they predict that response time should correspond to strength of preference, meaning that how much a decision-maker values one item over another ("value difference") should be expressed in response time, where smaller value differences correspond to longer response times (Echenique & Saito, 2017). This prediction has been supported in a variety of settings: risky choice (Alós-Ferrer & Garagnani, 2020; Konovalov & Krajbich, 2019), intertemporal choice (Amasino, Sul-

livan, Kranton, & Huettel, 2019; Dai & Busemeyer, 2014; Konovalov & Krajbich, 2019; Krajbich, Bartling, Hare, & Fehr, 2015), social decision-making (Frydman & Krajbich, 2018; Konovalov & Krajbich, 2019; Krajbich, Hare, Bartling, Morishima, & Fehr, 2015; Krajbich, Oud, & Fehr, 2014), and food choice (Clithero, 2018a, 2018b; Krajbich, Armel, & Rangel, 2010; Krajbich & Rangel, 2011; Milosavljevic, Malmaud, Huth, Koch, & Rangel, 2010; Towal, Mormann, & Koch, 2013).[1]

Previous work has demonstrated that people can empirically infer the preferences of others from their response times. The DDM specifically has been used as motivation for some of these studies, though the authors do not formally invert the DDM as we do. For example, in Konovalov and Krajbich (2020), participants in a strategic bargaining experiment inferred their opponents' preferences based on both explicitly stated and observed response times , and changed their behavior to make use of this knowledge. In another study, participants in Frydman and Krajbich (2018) played a strategic information cascade game, in which they had to determine the binary state of the environment, receiving both a probabilistic private signal about the environment and also observing other players' public choices. The DDM predicts that choices should be slow when players' private signals were in conflict with other players' public choices, which the authors observed. The study then examined how people behaved when presented with other players' response times. Participants used this information to infer how unsure other players were about their decisions: When public choices were incorrect, participants in the response-time condition were significantly more likely to follow their private signals compared to participants without access to other players' response times. Finally, an interesting consequence of the DDM's prediction that people will spend more time on harder decisions is that participants can act suboptimally under circumstances with a fixed time limit and payoffs allocated per-choice. Participants may spend a suboptimal amount of time on low-stakes choices with small value differences rather than high-stakes choices with large value differences. When researchers observed this result, they found that by forcing a cutoff time for all decisions they could significantly improve participants' payoffs (Krajbich et al., 2014; Oud et al., 2016).

Unlike previous work, we invert the DDM to predict how people will infer preferences from

[1] Previous work has studied response times in the framework of dual-process theory: If decisions are fast, this suggests that an intuitive process system is being used, and slow decisions suggest that a slow, deliberative process is being used (Rand, Greene, & Nowak, 2012; Rubinstein, 2007). Here, we interpret response time differences as representing differences in strength of preferences, on the basis of other work that shown that findings interpreted through the dual-process theory lens are consistent with single-process strength-of-preference predictions (Konovalov & Krajbich, 2020; Krajbich, Bartling, et al., 2015; Krajbich, Hare, et al., 2015; Zhao, Diederich, Trueblood, & Bhatia, 2019).

other decision-makers' choices and response times, fitting our model with participants' preference estimates about the decision-maker. This use of Bayes' rule to invert a decision-making model follows Lucas et al. (2014), Jara-Ettinger et al. (2016), Baker et al. (2017), and Jern et al. (2017), who inverted choice models to describe how people infer preferences from the outcomes others select. We continue to follow their tradition of using inverse decision-making models to explore theory of mind, extending this work to incorporate response times.

3.3 Bayesian inference over preferences from response times

We begin with a broad overview of our inverted Drift Diffusion Model (DDM), in which we use Bayesian inference to infer preferences from response times. The basic setup is that one person (the *observer*) watches a second person (the *decision-maker*) make a choice between two options. The observer then infers a distribution over how much the decision-maker likes each option (their *inferred utilities*) based on the option they chose and—critically—the amount of time they took to make the decision.

To make such an inference, the observer must have a *generative model* of the decision-maker's decision-making process. Here, we assume that this model is a DDM (Figure 3.1).[2] In this model, decisions are made on the basis of sequentially accumulated noisy evidence about the difference of the two options' utilities. The accumulation process runs until the total evidence in favor of one option or the other exceeds a threshold, at which point, the corresponding option is chosen. Formally, the evidence, x, is a dynamical system with dynamics [3]

$$dx = \mu + \sigma\, dW, \tag{3.1}$$

where μ is the *drift rate*, σ specifies the amount of noise in the integration, and W is a Weiner process (the continuous limit of a Gaussian random walk, also called Brownian motion). The evidence is initialized at zero. Following Milosavljevic et al. (2010), we assume that the drift rate is a linear function of the difference in values of the two items,

$$\mu = \beta(u_a - u_b), \tag{3.2}$$

where u_a and u_b are the utilities of the two items being chosen between, and the drift multiplier,

[3] We chose this model because it is commonly used in decision-making research and is computationally easy to work with. However, the framework can accommodate any model that defines a joint distribution over choices and response times.

β, can be interpreted as the decision-maker's sensitivity to differences in value. The accumulation process continues until the evidence crosses one of the decision boundaries. We assume symmetric constant boundaries, at a distance of θ from the initial point of zero. Intuitively, θ controls how careful the decision-maker is: larger values make decision slower but more accurate. If at any moment $x > \theta$, option a is chosen; we denote this event $a \succ b$. If $x < -\theta$, option b is chosen, that is, $b \succ a$. The time point at which this event occurs, t, is the response time (for simplicity, we do not consider the non-decision time variable that is often added to t to produce the response time). The DDM thus defines a probability distribution

$$p_{\text{DDM}}(a \succ b, t \mid u_a - u_b; \beta, \theta). \tag{3.3}$$

To infer the decision-maker's preferences given the observed choice and response time, the observer must invert this model. We can do this using Bayes rule, resulting in a posterior distribution over the utilities

$$p(u_a, u_b \mid a \succ b, t) \propto p(u_a)p(u_b) \cdot p_{\text{DDM}}(a \succ b, t \mid u_a - u_b; \beta, \theta), \tag{3.4}$$

where $p(u_a)$ and $p(u_b)$ capture the observer's prior distribution over utilities. For simplicity, we assume a standard Gaussian prior.

Equation 3.4 specifies how a rational agent should update their beliefs about another person's preferences based on an observed choice and response time. However, the inferences one draws depend on the parameters of the DDM: the drift multiplier, β, and the threshold, θ. Intuitively, these parameters can be thought of as individual difference or mental state variables that jointly determine the accuracy and speed of the decision. We explore the θ variable further in Experiment 3. For now, it is sufficient to note that the qualitative model predictions in Experiments 1 and 3 are insensitive to these parameters. However, for the purpose of plotting the predictions, we fit the parameters by minimizing the sum of squared errors between the model predictions and aggregate participant responses across the first two experiments. We treat the third experiment separately, as described in the methods for that experiment.

To summarize, the DDM serves as a generative model that relates decision-making parameters (the drift rate and threshold) to choices and response times. Critically, an observer can then invert this model to make inferences about the decision-maker's preferences. Here, we propose that human inferences from choices and response times will be consistent with such an ideal Bayesian observer model. To illustrate in a concrete example, recall our introductory example in which your friend takes longer to choose "movie" when it is paired with science museum than when it is paired

with hiking. In considering her responses, you would conclude not only that she really wants to see a movie, but that she seems to like science museums more than hiking. Remarkably, this intuitive inference falls naturally out of inverting the DDM: Her longer response time in the movie-museum pairing reflects a more shallow drift rate, which results from the movie-museum utility difference being smaller than the movie-hiking difference. Since the movie has the highest utility (after all, she chose it twice), this means that the museum has higher utility than hiking.

In the remainder of the paper, we empirically test three predictions of our model. Experiment 1 tests the prediction that observing faster choices indicates that the decision-maker has a larger preference difference. Our inverse DDM predicts this because a large difference in utilities between two items will result in a steeper drift rate, which leads to a faster decision. In Experiment 2, we test the prediction that people can draw inferences about novel choice pairs even when the choices alone provide no information about the relative utility. In particular, we present participants with scenarios similar to the movie/hiking/museum example and test whether they integrate information across observations consistent with inverting the DDM. Finally, in Experiments 3, 4A, and 4B, we examine whether people's conclusions about preferences from response times is affected by background knowledge about the decision-maker's mental state, specifically whether the decision-maker is feeling cautious or careless. This corresponds to performing inference about the decision-maker's drift rate (i.e., the relative utilities of the options) conditioned on different decision thresholds.

3.4 Experiment 1: Inferring preferences from response times

We test the core prediction of the DDM that decisions will be faster when the difference in preference for the two items is larger. In the model, a larger preference difference results in a higher drift rate and more rapid movement towards a decision boundary. Thus, if people employ a model like the DDM when inferring other's preferences from their response times, they should rate a chosen item as being more strongly preferred when the decision is made more quickly.

3.4.1 Methods

Participants

481 participants with U.S. IP addresses were recruited through Amazon Mechanical Turk. Participants were paid $1.00 in compensation. Participants were excluded from the study if, for the critical questions, they indicated the decision-maker preferred a non-chosen item more than once, indicating lack of attention. 66 participants were excluded for a total of 415 participants.

We used power analyses to determine the sample size for Experiments 1, 2, and 3 based on effect sizes observed in pilot studies. These power analyses, our exclusion criteria, our stimuli, and our planned procedures and analyses were preregistered here: https://osf.io/8n3kd) Participants could only complete one of the pilot experiments or main experiments.

Experiments 1, 2, and 3 were IRB-approved by University of California, Berkeley, Committee for Protection of Human Subjects / Office for Protection of Human Subjects, Protocol ID: 2015-05-7551, Protocol Title: Cognitive Research Using Amazon Mechanical Turk (Expedited), with all participants giving informed consent.

Stimuli

In all experiments, participants viewed surveys created on Qualtrics containing text and videos. Experiments 1, 2, and 3 used the same set of videos (or a subset of these videos), which were created by filming a decision-maker making eight choices. The videos began by the decision-maker pulling two pieces of paper apart, which were covering two labeled bowls (e.g. A and B; bowls could be labeled A, B, or C). The decision-maker thought about their options while staring between the two bowls, and then they reached into one of the bowls to make a choice. The decision-maker made choices after 3, 5, 7, and 9 seconds after video onset, and made choices with either their left or right hand, for a total of eight videos. To counterbalance asymmetries in the videos, the number of videos used was doubled by creating flipped copies over the vertical axis, for a total of 16 videos. To counterbalance the effects of A, B, and C labels, the labels of the 16 videos were edited to use the label combinations AB, BA, CA, AC, BC, and CB, for a total of 96 videos. The text was modified to match the video labels. For narrative simplicity, we will only refer to the AB label combinations. Stimuli for all experiments are available on OSF: https://osf.io/pczb3.

Procedure

After consenting, participants were shown introductory text: "In this experiment, you will see videos of someone choosing between an item in one bowl and an item in another bowl. There are 8 videos. Please do your best to answer the questions afterward."

Participants then read: "Two items, A and B, are inside the bowls below. This person has just been asked to choose which item they want. Please watch the video." Participants then watched a video in which a decision-maker chose an item from bowl A or B within 3, 5, 7, or 9 seconds from the onset of the video. After watching the video, participants were asked: "What are the person's feelings about item A and item B?" Participants then moved a slider containing values from

0 (labeled "Strongly Prefers A") to 100 (labeled "Strongly Prefers B") and set initially at 50 (labeled "Neutral Between A and B"). No grid lines were shown, nor the numbers 0-100, and participants were required to click and move the slider (but they could move it back to 50 if desired). After answering this question the participant could advance to the next trial, which displayed on another page (Figure 3.2). Participants could replay videos as often as they wished, and all of the text/video/questions were on the same page. Participants could not return to a previous page. There were eight trials: Participants saw a total of eight videos (A/B choice $\times$ 3/5/7/9 seconds) in random order.

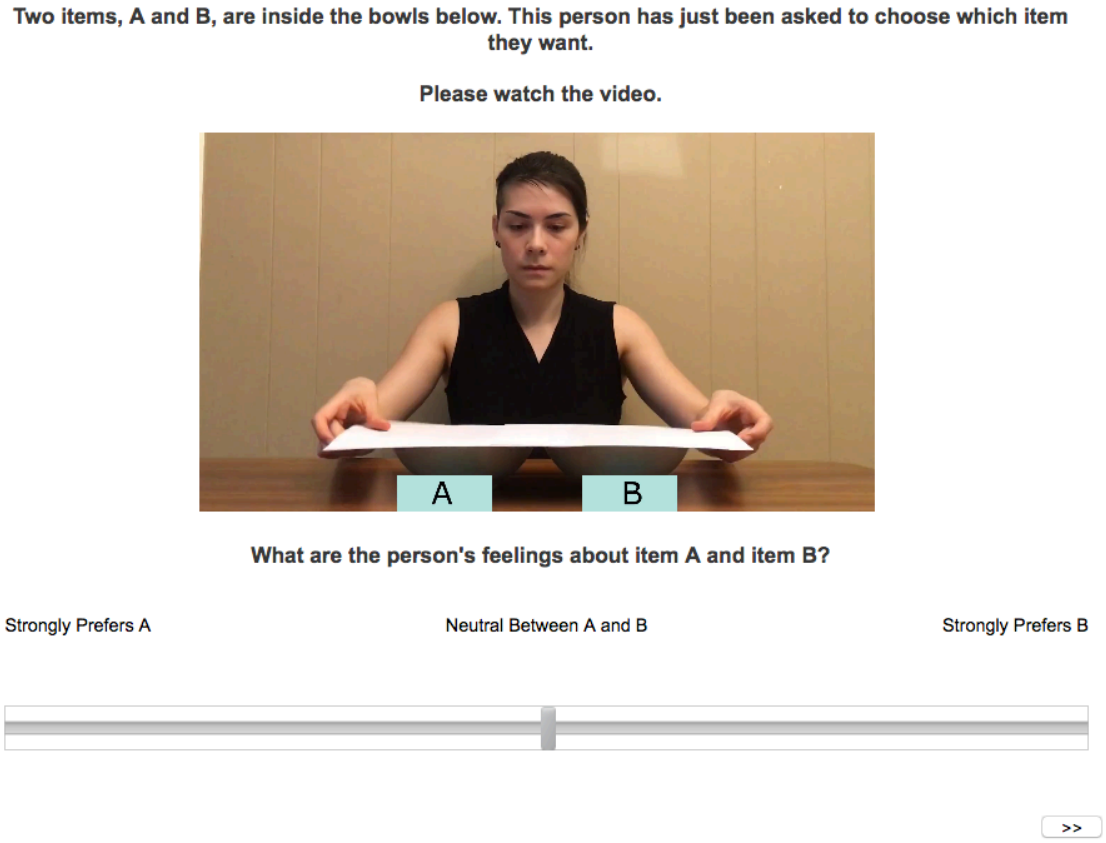


Figure 3.2: Experiment 1 Stimulus. Participants watched a 3, 5, 7, or 9 second video of the participants choosing between A or B, then answered the question below.

We will refer to the choices as between A and B for narrative simplicity. However, participants were randomly assigned to one of 12 counter-balancing variants in which the labels and video orientations were randomized. Within every variant, the labels were held constant (options: AB, BA, AC,

CA, BC, CB) and the video orientation was held constant (the videos were either as filmed, or were flipped over the vertical axis).

After watching all eight videos and answering the accompanying questions, participants had the opportunity to write comments then exited the survey.

Model

We assume that participant responses are based on the posterior mean estimate of the difference in utilities of the two items. This is defined as

$$\mathbb{E}[u_a - u_b \mid a \succ b, t] = \int\int (u_a - u_b) \cdot p(u_a, u_b \mid a \succ b, t)\, du_a du_b. \tag{3.5}$$

However, because the DDM likelihood depends only on the difference in values (Equation 3.3), we can re-express Equation 3.5 as a single integral over the difference,

$$\mathbb{E}[\delta_u \mid a \succ b, t] = \int \delta_u\, p(\delta_u \mid a \succ b, t)\, d\delta_u, \tag{3.6}$$

where $\delta_u = u_a - u_b$ and

$$p(\delta_u \mid a \succ b, t) = \frac{1}{Z}\text{Normal}\left(\delta_u; \mu = 0, \sigma^2 = 2\right) \cdot p_{\text{DDM}}(a \succ b, t \mid \delta_u). \tag{3.7}$$

Note that the variance of the difference between two Gaussians is the sum of the variance of each, hence $\sigma^2 = 2$. We have suppressed β and θ for concision. We compute the normalizing constant, Z, and the integral in Equation 3.6 by adaptive numerical integration (Genz & Malik, 1980).

To convert this utility difference (which is in arbitrary units) to the scale defined by the slider participants used to make a response, we use a scaling parameter, such that the predicted response is $\alpha\, \mathbb{E}[u_a - u_b \mid a \succ b, t]$. We fit α (along with β and θ) to aggregate participant behavior by minimizing the sum of squared errors.

As one would expect, the DDM predicts that a choice of item a over item b is increasingly likely as $u_a - u_b$ increases. However, the choice alone only weakly constrains the size of the difference. As illustrated in Figure 3.1, the response time, t, provides much more information. With a strong preference, a fast response is quite likely and a slow response is quite unlikely. With a weak preference, probability is more evenly spread across response times. As a result, when we invert the DDM, we find that a strong preference is more likely given a fast response (because a fast response is likely given a strong preference) and a weak preference is more likely given a slow response (because a slow

response is unlikely given a strong preference).

3.4.2 Results

In Experiment 1, we showed participants videos in which a decision-maker took 3, 5, 7, or 9 seconds to decide between two items, then asked participants about the decision-maker's preferences about the items. We expected that participants would infer that the decision-maker preferred the chosen item more when the response time was shorter (e.g. 3 seconds) compared to when the response time was longer (e.g. 9 seconds), as predicted by our model. This is indeed what we observed, with a qualitatively close match between the model's predictions and the results (Figure 3.3).

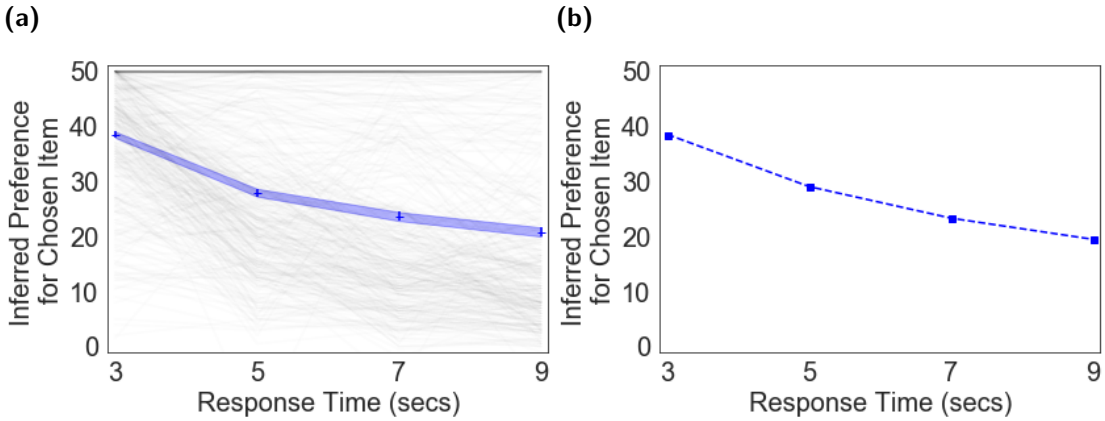


Figure 3.3: Experiment 1 Results. (a) Experimental Results. The mean ± SE of inferred preference for the chosen item, n=415, is shown for each of the response time conditions. Responses were reoriented to be 0-50: 0 indicates the participant felt the decision-maker was neutral between items A and B, and 50 indicates the participant felt the decision-maker strongly preferred the chosen item (A or B, whichever was appropriate to the trial). Light grey lines indicate individual participants' averaged values for each time point. (b) Model Predictions.

To analyze the effect of response time, we created a linear mixed effects model with response time as a continuous fixed effect, subject as a random effect, and participants' preference judgments (0-100, 50 was "Neutral Between A and B," 0 was "Strongly Prefers A" and 100 was "Strongly Prefers B") as the continuous output variable. Each participant had eight data points: two answers each for the 3, 5, 7, and 9-second questions. The estimate (beta parameter) for the fixed effect of response time was -2.87 (95% CI=[-3.03,-2.71], t(2904.0)=-35.16, p<2e-16, t-tests using Satterthwaite's method, η_p^2=0.30). (The estimate for the fixed effect intercept was 44.91, 95% CI=[43.34,46.48], t(1003.4)=56.10, p<2e-16, t-tests using Satterthwaite's method, η_p^2=0.76.) The estimated power of

the response time predictor for α=0.01 was 100.0% (95% CI=[98.17,100.0]) using a Type-III *F*-test from the R package "car," generated via 200 simulations with the R package "simr."

3.4.3 Discussion

As predicted by our model, participants inferred that a decision-maker preferred an item more the less time they spent selecting the item. This effect was statistically significant, and qualitatively matched our model results. Interestingly, the individual participant responses in Figure 3.3 suggest that the main effect of response time may be even stronger than indicated by the mean responses, because a minority of participants (40/415 = 9.6%) do not take the decision-maker's response time into account. Instead, they inferred that the decision-maker had the maximal preference for the chosen item on every question (shown in Figure 3.3 by a set of {50, 50, 50, 50} responses, indicating the participant moved the slider to 100 on four trials and 0 on four trials).

3.5 Experiment 2: Predicting novel choices based on relative strength of preferences

In Experiment 1, we demonstrated that participants could infer preferences of a decision-maker from their choices and their response times. However, people regularly perform more complex inferences than that: If your friend did not hesitate at all before choosing movies over hiking, but hesitated a while before choosing movies over science museums, then you could guess that she preferred science museums to hiking. We now explore this ability to predict a decision-maker's choice in a novel comparison, based on having inferred their relative strength of preferences from response times in previous choices.

Inferring graded preferences from choices (or response times) that will allow prediction for novel comparisons is not often studied. The most closely related literature surrounds *transitive inference*. In transitive inference paradigms, participants—children, adults or animals—are asked to draw conclusions like if B > A, and A > C, then B > C. In these paradigms, associations are often manually taught between previously-unassociated items or assertions, and transitive inference is solely a function of choices (e.g. Acuna, Sanes, & Donoghue, 2002; Harris & McGonigle, 1994; Maybery, Bain, & Halford, 1986; Thayer & Collyer, 1978). It is rare to show participants executing transitive inference about preferences, which are distinct in that they are often socially learned through implicit cues in addition to appearing probabilistic and non-causal. Hu, Lucas, Griffiths, and Xu (2015) is an exception, demonstrating that young children can perform indirect, graded transitive inference of another decision-maker's preferences based on observing their choices. Children in this study

inferred that a puppet preferred object B over C (B > C), after watching the puppet choose B consistently over A (B $\gg$ A) and the puppet choosing C somewhat consistently over A (C > A).

Similar to Hu et al. (2015), we test whether people can infer another decision-maker's graded preferences by observing their choices. However, rather than watching the decision-maker make choices repeatedly, we had participants watch the decision-maker make two decisions. We investigate whether people can infer a decision-maker's preferences enough to predict their selection on a novel choice after only having observed the decision-maker's choices and response times twice. We can do this because it can be much more efficient to use response time and choice information together than choice information alone. If A were chosen over B, and A were chosen over C, people would have to observe more choices to statistically infer the graded preferences that would allow them to make a prediction for an unseen choice between B and C. Using response time information, people could make this prediction immediately, given they had inferred the decision-maker's relative strength of preferences after observing their previous choices and response times.

In Experiment 2, participants saw pairs of videos. In the first video, the decision-maker made a choice between two items, A and B, within 3 seconds. In the second video, the decision-maker made a choice between the original item and another item, A and C, within 9 seconds. Participants were then told that the decision-maker had to make a choice between B and C, and were asked what choice they thought the decision-maker would make and how likely that choice was. Our model predicts that participants would anticipate the decision-maker's selection on the novel choice, having inferred the decision-maker's preferences based on response times and choices from the previous two decisions. Our model also proposes that participants' inferred likelihoods of the predicted choice would be lower than in the control condition in which participants only needed to use choices to make the prediction (and did not need to take response time into account).

3.5.1 Methods

Participants

478 participants with U.S. IP addresses were recruited through Amazon Mechanical Turk. Participants were paid $1.50 in compensation. No exclusion criteria were applied. The sample size was determined by a power analysis based on the effect size of pilot studies and was preregistered.

Stimuli and Procedure

Video stimuli were the same as in Experiment 1.

After consenting, participants read introductory text: "In this experiment, you will see videos of someone choosing between an item in one bowl and an item in another bowl. There are 8 sets of videos. Please do your best to answer the questions afterward." Participants then entered the main experiment.

In Experiment 2, participants saw pairs of videos. Participants saw the text "This person is choosing between two items, A and B," and then were shown a first video in which the decision-maker makes a choice within 3 seconds. On the same page, participants were then presented with the text "Now they are choosing between A and C," and shown another video of a decision-maker making a choice within 9 seconds. The participant was then shown, also on the same page, "This person is now offered a choice between items B and C. What choice do you think they'd make, and how likely do you think it is that they'd make that choice?" Participants then moved a slider containing values from 0 (labeled "Very likely B") to 100 (labeled "Very likely C") and set initially at 50 ("Equally likely"). No grid lines were shown, nor the numbers 0-100, and participants were required to click and move the slider (but they could move it back to 50 if desired). After watching the two videos and answering
the question, participants could advance to the next page, where they watched another pair of videos and answered another question (Figure 3.4). There were eight pages/questions total. Participants could replay videos as often as they wished, and all of the text/videos/questions were on the same page. Participants could not return to a previous page. After watching all eight pairs of videos and answering the accompanying questions, participants had the opportunity to write comments then exited the survey.

For narrative simplicity, we will refer to the first video as a 3-second choice between A and B, and the second video as a 9-second choice between A and C. However, randomized counter-balancing variants were used in the experiment. We randomized the order of which video was presented first within-participant (half of the trials showed the 3-second video first, and half of the trials showed the 9-second video first). Between-participants, we randomized the order of the final choice between participants (in one variant, participants made a choice between B and C, and in another variant participants made a choice between C and B), the labels (one label was kept consistent on one side, resulting in the following variants: AB vs AC, BA vs CA, BA vs BC, AB vs CB, CA vs CB, and AC vs BC), and video orientation (half of the variants had all videos flipped over the y-axis). This resulted in a total of 24 variants (2 final choice options $\times$ 6 label pairs $\times$ 2 video orientations), and 8 questions per participant (2 orderings of 3/9-second videos $\times$ 4 choice options (R/R, R/L, L/R, R/R, where "R/L" represents the decision-maker choosing the right-most object in the first video and the left-most object in the second video)).

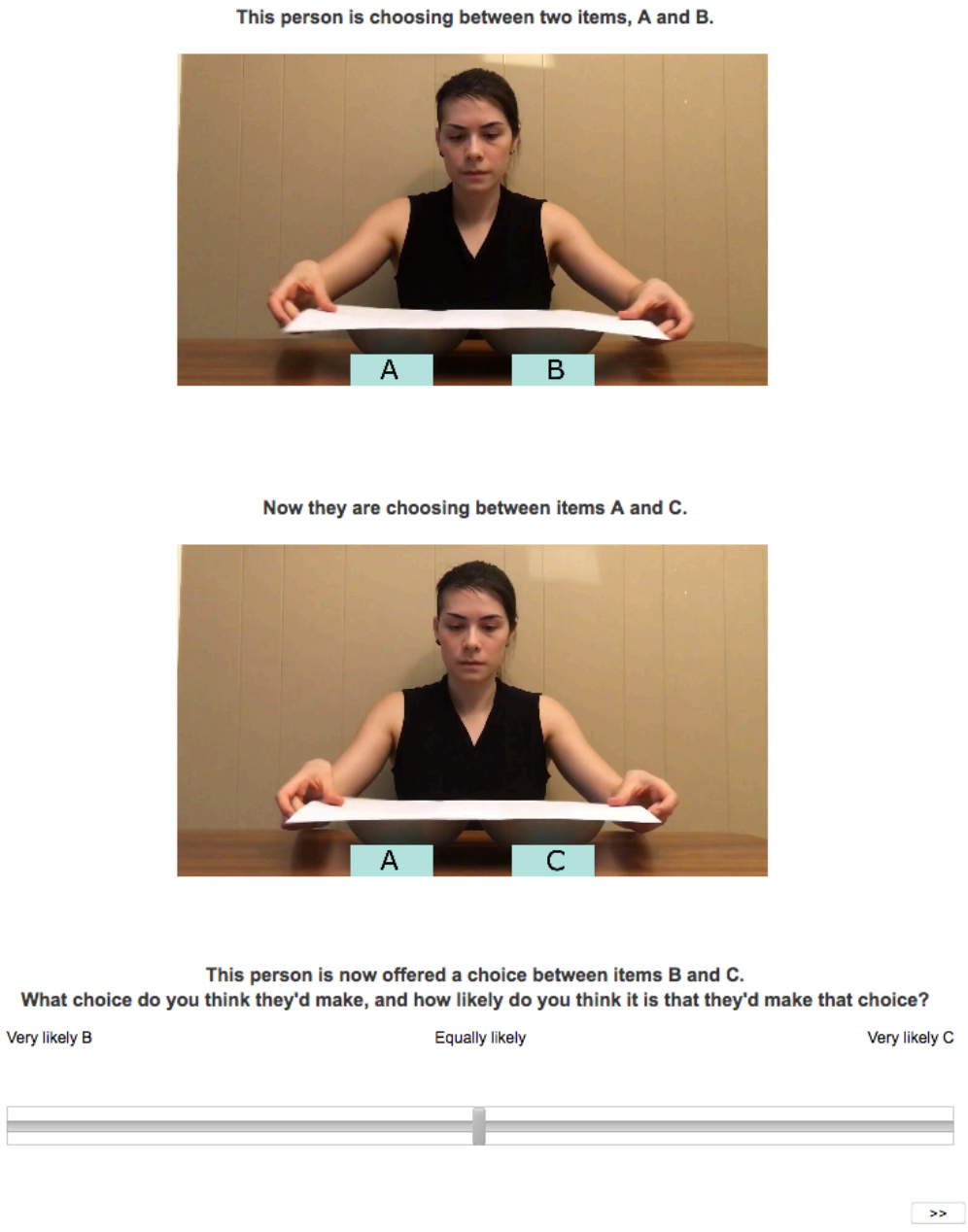


Figure 3.4: Experiment 2 Stimulus. Participants watched a 3-second video of the decision-maker deciding between A and B, and a 9-second video of the decision-maker deciding between A and C. Participants then indicated whether they believed the decision-maker would be most likely to choose B or C in a novel choice.

Participants were randomly assigned to one of these 24 variants, within which they viewed the 8 pairs of videos wherein labels and video orientation were held constant but response time and choices varied. A given participant would be shown the video pairs in random order, but the videos within the pairs were ordered.

Analysis

Our stimuli resulted in four conditions. Participants watched the decision-maker choose between A and B in one choice, then A and C in another, and in the critical question participants were asked whether the decision-maker would prefer B or C in a novel choice. We thus defined the four conditions based on whether the decision-maker chose B or C in either of the observed choices. The condition "NeitherChosen" indicated that the decision-maker never chose B or C in the observed choices (choosing A instead), and the condition "BothChosen" indicated that the decision-maker chose B in one choice and C in another. The condition "FastChoice" indicated that the decision-maker chose B once, and "Slow choice" indicated that the decision-maker chose C once (Figure 3.5). Each participant answered two questions for each of the four conditions, for a total of eight data points.

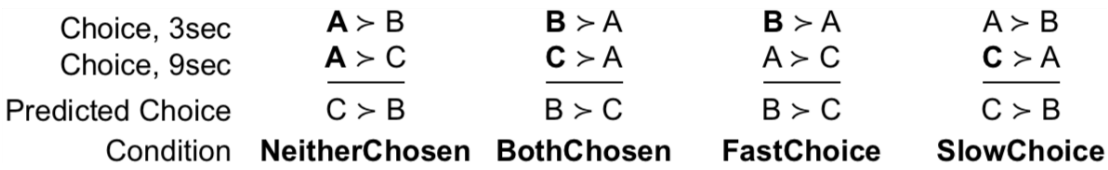


Figure 3.5: Experiment 2 Conditions. Participants watched a 3-second video of the decision-maker deciding between A and B, and a 9-second video of the decision-maker deciding between A and C. In the critical question, participants indicated whether they believed the decision-maker would be more likely to choose B or C in a novel choice. The model's predictions for participants' predicted choices are shown. Four conditions were defined based on whether the decision-maker chose B or C in either of their choices (3-second and 9-second). In the "NeitherChosen" condition, the decision-maker chose neither B nor C, and in the "BothChosen" condition, the decision-maker chose B and C. For the non-inference control conditions, in the "FastChoice" condition, the decision-maker chose B quickly, and in the "SlowChoice" condition, the decision-maker chose C slowly.

The first two conditions, "NeitherChosen" and "BothChosen," require that the participant make an inference based on response time, since in the "NeitherChosen" condition the decision-maker never preferred B or C in the observed choices, and in the "BothChosen" condition the decision-maker preferred both. Thus when the participant was asked whether the decision-maker preferred B or C, the answer was ambiguous based solely on the choices the participant had seen.

The second two conditions, "FastChoice" and "SlowChoice," did not require that the participant make an inference based on response times, just choices, since the decision-maker preferred either B or C in the choices they made in the videos. Thus when the participant was asked whether the decision-maker preferred B or C, the decision-maker should logically choose whichever choice the decision-maker preferred in one of the videos.

Model

We assume that participants respond with the posterior mean probability of the unseen choice given the two observed choices and response times. To compute this value, we first compute a posterior over each item's value, and then integrate over these values to produce a predicted choice. Here, we specify these steps for the "NeitherChosen" condition; the derivations for the other conditions have the same form. Let $\mathcal{D} = \{a \succ b, t_{ab}, a \succ c, t_{ac}\}$ denote the observed data. The posterior over values is then given by

$$p(u_a, u_b, u_c \mid \mathcal{D}) \propto \varphi(u_a)\varphi(u_b)\varphi(u_c) \cdot p_{\text{DDM}}(a \succ b, t_{ab} \mid u_a - u_b) \cdot p_{\text{DDM}}(a \succ c, t_{ac} \mid u_a - u_c), \quad (3.8)$$

where φ is the standard normal pdf, and the posterior over choice is given by marginalizing over the three values as well as the predicted decision time,

$$p(b \succ c \mid \mathcal{D}) = \int\int\int\int_0^\infty p_{\text{DDM}}(b \succ c, t_{bc} \mid u_b - u_c) \cdot p(u_a, u_b, u_c \mid \mathcal{D})\, dt_{bc}\, da\, db\, dc. \quad (3.9)$$

At a high level, the key model prediction is that people will be able to predict a novel choice even when the previous choices (without response times) do not provide any information about which option is preferred. Consider the "NeitherChosen" case. Here, the observer must predict a choice between B and C after observing A being chosen over both of them. This implies that $u_a > u_b$ and $u_a > u_c$, but provides no information about $u_b - u_c$. However, considering the response times of each choice (short and long, respectively) provides additional information about the relative strength of preference in the first two choices. Following the predictions of Experiment 1, the observer infers that $u_a - u_b$ is large because this decision was made quickly and that $u_a - u_c$ is small because this decision was made slowly. Thus, u_c must be greater than u_b, and C is more likely to be chosen over B. This line of reasoning can be summarized in notation as:

$$a \succ b \wedge a \succ c \wedge t_{ac} > t_{ab} \implies u_a - u_c < u_a - u_b \implies u_c > u_b \implies p(c \succ b) > \frac{1}{2}. \quad (3.10)$$

The logic is similar in the "BothChosen" case, but inverted:

$$b \succ a \wedge c \succ a \wedge t_{ac} > t_{ab} \implies u_c - u_a < u_b - u_a \implies u_c < u_b \implies p(b \succ c) > \frac{1}{2}. \quad (3.11)$$

For comparison, we also consider cases where the previous choices alone *do* constrain the prediction about the novel choice, by transitivity of preference. In the "FastChoice" condition we have

$$b \succ a \wedge a \succ c \implies u_b > u_a \wedge u_a > u_c \implies u_b > u_c \implies p(b \succ c) > \frac{1}{2}, \quad (3.12)$$

and in the "SlowChoice" condition,

$$a \succ b \wedge c \succ a \implies u_a > u_b \wedge u_c > u_a \implies u_c > u_b \implies p(c \succ b) > \frac{1}{2}. \quad (3.13)$$

The model makes a stronger prediction in the "FastChoice" and "SlowChoice" conditions because choices provide stronger and more direct evidence of preference ordering compared to differences in response time.

3.5.2 Results

Our model predicts that participants would anticipate a decision-maker's novel choice based on inferring the relative strength of that decision-maker's preferences from previous response times and choices. Specifically, participants should predict the decision-maker's choice better than chance. Moreover, participants' inferred likelihoods of the predicted choice should be lower than the control conditions in which only choices (and not response times) were necessary. We compare our results qualitatively to these model predictions (Figure 3.6).

Difference from chance, all conditions

To determine whether participants could make predictions for a novel choice better than chance, we conducted the equivalent of one-sample *t*-tests for a linear mixed effects model.[4] Condition was a categorical fixed effect ("NeitherChosen"/"BothChosen"/"FastChoice"/"SlowChoice"), subject was a random effect, and likelihood judgment was a continuous output variable.

[4]Specifically, we translated participants' likelihood judgments to be between -50 and 50 (where 0 represented a judgement that it was equally likely the decision-maker would choose either of the two items) then did not fit the fixed effect intercept (which defaults to 0).

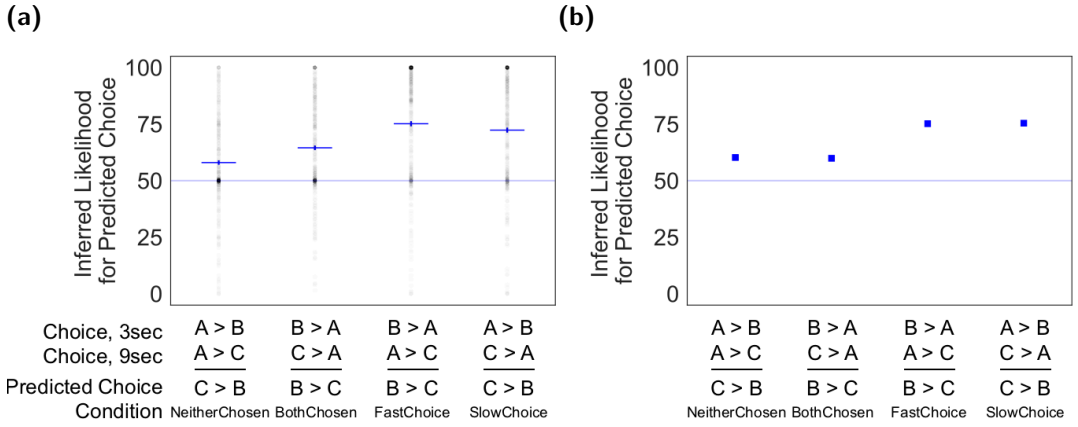


Figure 3.6: Experiment 2 Results. (a) Experimental Results, Mean ± SE of Inferred Likelihood Estimates for the Predicted Choice, n=478. Shown in blue are participants' averaged responses for each of the conditions; the line at 50 indicates equal likelihood between choices (chance); black points indicate individual participant averaged responses for each condition. Participants saw the decision-maker choose between A and B, then A and C, and had to infer whether the decision-maker preferred B or C in the the unseen choice (the model's predicted choices are shown). The "NeitherChosen" and "BothChosen" conditions required that participants infer the decision-maker's likelihood of preferring an item from choices and response time, while the "FastChoice" and "SlowChoice" conditions only required that participants infer from the decision-maker's choices. (b) Model Predictions.

In the "NeitherChosen" condition, we predicted better-than-chance performance: specifically that if the decision-maker chose A $\succ$ B quickly, and A $\succ$ C more slowly, that participants would infer that the decision-maker had preferences A > C > B. From our linear mixed effects model described above, the estimate (beta parameter) for the "NeitherChosen" condition was 7.99 (95% CI=[5.89,10.10], t(1681.5)=7.46, p=1.4e-13, t-tests using Satterthwaite's method, η_p^2=0.03),[5] in the expected direction C > B. The estimated power of the "NeitherChosen" condition against chance for α=0.01 was 100.0% (95% CI=[98.17,100.0]) using a t-test, generated via 200 simulations with the R package "simr."

In the "BothChosen" condition, we expected that if the decision-maker chose B $\succ$ A quickly, and C $\succ$ A slowly, then participants would infer that the decision-maker had preferences B > C > A above chance. From our linear mixed effects model described above, the estimate (beta parameter) for the "BothChosen" condition was 14.53 (95% CI=[12.43,16.63], t(1681.5)=13.55, p<2e-16, t-tests using Satterthwaite's method, η_p^2=0.10), in the expected direction B > C. The estimated power

[5] All η_p^2 effect sizes were calculated using the assumption that $t^2 = F$, followed by $\eta_p^2 = F * df_{num}/(F * df_{num} + df_{res})$ (e.g. Lakens, 2013) with $df_{num} = 1$.

of the "BothChosen" condition against chance for α=0.01 was 100.0% (95% CI=[98.17,100.0]) using a *t*-test, generated via 200 simulations with the R package "simr." These results match our model predictions, with one difference being that the model predicted a more similar level of inference for the "NeitherChosen" and "BothChosen" conditions than was observed empirically.

We also expected that participants would make the appropriate choices in the "FastChoice" and "SlowChoice" control conditions, that they would do so above chance performance, and that they would do so more accurately than in the "NeitherChosen" and "BothChosen" inference conditions. The "FastChoice" and "SlowChoice" conditions act as control conditions: In these conditions, participants only needed to use choice information to correctly infer the decision-maker's preferences, instead of also needing response times. From our linear mixed effects model, the estimate (beta parameter) for the "FastChoice" condition was 25.17 (95% CI=[23.07,27.27], t(1681.5)=23.47, p<2e-16, *t*-tests using Satterthwaite's method, η_p^2=0.25), in the expected direction B > C. The estimated power of the "FastChoice" condition against chance for α=0.01 was 100.0% (95% CI=[98.17,100.0]) using a *t*-test, generated via 200 simulations with the R package "simr." The estimate for the "SlowChoice" condition was 22.34 (95% CI=[20.24,24.44], t(1681.5)=20.84, p<2e-16, *t*-tests using Satterthwaite's method, η_p^2=0.21), in the expected direction C > B. The estimated power for α=0.01 was 100.0% (95% CI=[98.17,100.0]) using a *t*-test. These results matched our model predictions. One difference was that the model predicted slightly more similar level of inference across these conditions than was observed empirically.

The "FastChoice" and "SlowChoice" results serve as an baseline for accuracy based purely on choice compared to the "NeitherChosen" and "BothChosen" results, which required inferences based both on choice and response time. We explore this comparison directly next.

Comparison between conditions: timing-inference conditions vs choice control conditions

In addition to examining how our results differ from chance, we also examine how the timing-based inference conditions ("NeitherChosen"/"BothChosen") compare to the control conditions ("FastChoice"/"SlowChoice"). To this end, we constructed a linear mixed effect model where condition was a categorical fixed effect, subject was a random effect, and participants' likelihood judgments were a continuous output variable. However, unlike the previous model, the conditions were merged ("NeitherChosen/"BothChosen" and "FastChoice"/"SlowChoice") to create two comparison groups, and these groups were compared to each other.

The estimate (beta parameter) for the fixed effect of group (timing-inference conditions versus

control conditions) was -12.49 (95% CI=[-14.22,-10.77], t(3345.0)=-14.17, p<2e-16, t-tests using Satterthwaite's method, η_p^2=0.06). (The estimate for the fixed effect intercept was 73.76, 95% CI=[72.04,75.47], t(842.1)=84.24, p<2e-16, t-tests using Satterthwaite's method, η_p^2=0.89.) The estimated power of the group predictor for α=0.01 was 100.0% (95% CI=[98.17,100.0]) using a Type-III F-test from the R package "car," generated via 200 simulations with the R package "simr." Our empirical results matched model results, showing a statistically significant pairing between the timing-based inference conditions and the control conditions, with the control conditions showing higher inferred likelihoods for predicted choice.

3.5.3 Discussion

In Experiment 2, we asked whether participants could make a prediction about an unseen choice based on inferring the relative strength of a decision-maker's preferences from response time and choice information from two observed choices. Our model results and empirical results were in close alignment: In all conditions, participants inferred the predicted choice above chance. Moreover, participants had lower inferred likelihoods for the predicted choice for the conditions in which they needed to use both response time and choice information (the "NeitherChosen" and "BothChosen" inference conditions) compared to when they only needed to use choice information (the "FastChoice" and "SlowChoice" conditions, which serve as an baseline for accuracy based purely on choice).

Specifically, in the "NeitherChosen" condition, participants saw the choice A $\succ$ B made quickly, the choice A $\succ$ C made slowly, and the model predicted that they would infer that C > B, based on the DDM reasoning that items closer in value take a longer time to decide between. In the "BothChosen" condition, participants saw the choice B $\succ$ A made quickly, the choice C $\succ$ A made slowly, and the model predicted that they would infer that B > C, again based on the reasoning that items closer in value take a longer time to decide between. Interestingly, participants were better able to perform the inference for the "BothChosen" condition compared to the "NeitherChosen" condition, while the model predicted very similar inferred likelihoods for both of these "inference" conditions. Intuitively, it can feel easier to choose whether B or C was better if either of these items were chosen compared to if neither were chosen, but this intuition was not captured by the model. Future versions of the model should capture this more subtle behavior, perhaps by incorporating a bias for reduced timing for chosen items.

In the non-inference control conditions, participants inferred the predicted choice similarly across the two conditions "FastChoice" (the participants saw B $\succ$ A chosen quickly, A $\succ$ C cho-

sen slowly, and had to infer B $\succ$ C) and "SlowChoice" (the participants saw A $\succ$ B chosen quickly, C $\succ$ A chosen slowly, and had to infer C $\succ$ B), as predicted by the model. Participants did have slightly higher inferred likelihoods for the predicted choice on the "FastChoice" condition, in which the target item was chosen quickly rather than slowly, which was not represented in the model. This was a relatively minor effect, but future work could to explore whether this element of human psychology could be captured in a modified DDM or with a different set of well-justified parameters. Relevantly, the model predictions for Experiment 2 were highly varied based on the parameter settings, so it could be interesting to explore what kind of behavior is described under those different settings.

3.6 Experiment 3: Sensitivity to a decision-maker's mental state

We next investigated whether people would incorporate a decision-maker's mental state when reasoning about response times and implied preferences. If a decision-maker quickly chose between two items, it could have been because one of the options was obviously preferred. However, if the decision-maker was feeling careless rather than cautious that day, their quick decision-making could also be attributed to being tired and rushing through decisions. Both factors—the value difference between the two choices, and the decision-maker's overall carefulness—could influence response time. We asked whether participants would incorporate the decision-maker's carefulness in reasoning about their preferences.

From a modeling perspective, in the DDM, a decision-maker's carefulness is naturally modeled by the decision threshold θ, which determines the amount of evidence that must be accumulated in favor of an item before a decision is made. The higher the threshold, the less sensitive the decision-maker will be to random fluctuations in the evidence, and thus the less likely they will be to make an error by selecting the item that they actually disprefer. However, this robustness comes at a cost: Accumulating evidence takes time, and so a decision-maker with a high threshold will make slower decisions. Importantly, the decision time depends on both the drift rate and the threshold. A slow decision can result from either a low drift rate (small preference difference) or a high threshold (high caution) (Figure 3.7).

The combined influence of drift rate and threshold on decision time results in a critical prediction of our model: knowing a decision-maker's carefulness should affect one's inferences about their preferences. Specifically, if you believe that a decision-maker is very cautious (has a high threshold), and you observe them make a fast decision between two items, you should infer a strong preference for the chosen item, because a high threshold is unlikely to be reached quickly unless the evidence is

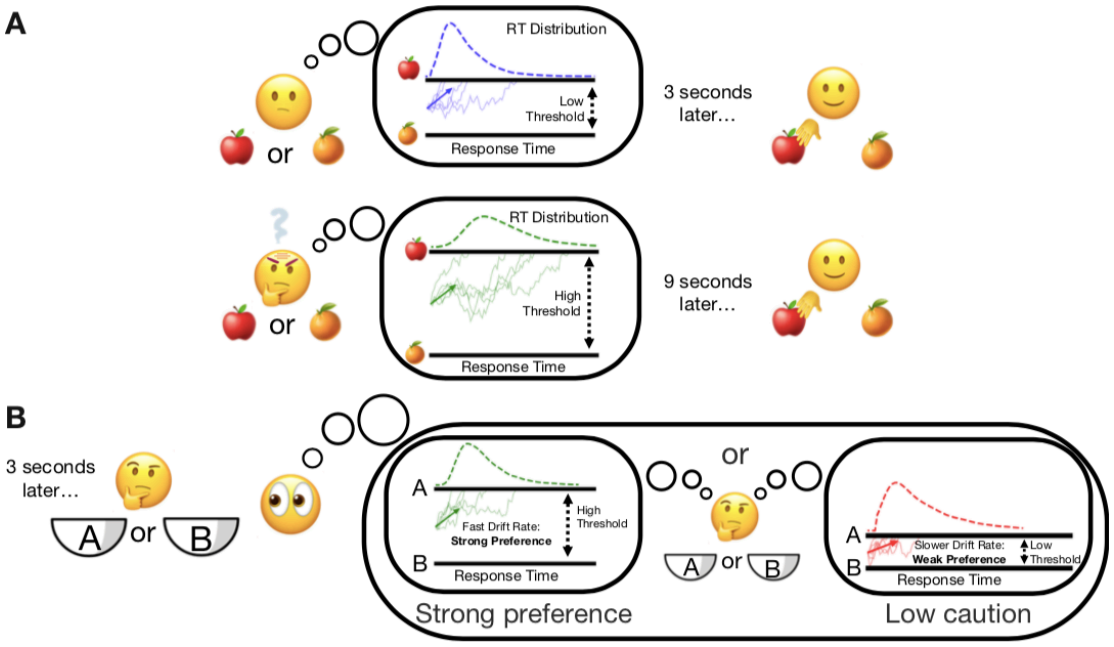


Figure 3.7: Schematic describing the DDM's threshold parameter, and the inference task in Experiment 3. (A) The DDM's parameter "threshold" (θ) describes how careful a decision-maker is. If a decision-maker has a fixed drift rate (meaning that their perceived value difference or preference for one item over the other is the same), a more cautious decision-maker (having a high threshold) will take longer to make a choice than a less cautious decision-maker (having a low threshold). (B) In an inverted DDM, if the observer does not know the decision-maker's drift rate (preferences) or threshold (carefulness), the observer must infer whether a decision was fast because the decision-maker had a strong preference for that item, or because the decision-maker was being careless and not thinking too much about their decision. In Experiment 3, participants are told the decision-maker's carefulness and response time and must infer their strength of preference.

strong. In contrast, if you believe the decision-maker to be careless (low threshold), a fast decision is consistent with moderate or even no preference difference.

In Experiment 3, participants watched decision-makers making decisions quickly (3 seconds) or slowly (9 seconds), having been described as cautious (high threshold) or careless (low threshold). Participants were then asked how much they believed the decision-maker valued their chosen item. We hypothesized based on model predictions that if a decision-maker made a choice very quickly and was described as being cautious, then participants would infer that the decision-maker valued one of the items a lot more than the other. If a decision-maker made a choice very quickly but was described as careless, we expected that participants would attribute this quickness only partly to how much the decision-maker valued the items, and partly to the decision-maker's mental state. We ex-

pected the same pattern of results if a decision-maker made a choice slowly. Finally, we expected that there would be an interaction effect between response time and carefulness/threshold. Specifically, we predicted that participants would learn more about a decision-maker's preferences if they were cautious (where a fast decision versus a slow decision would be more meaningful) compared to if they were careless (where a fast decision versus a slow decision would be less meaningful or indicative of their preferences).

3.6.1 Methods

Participants

481 participants with U.S. IP addresses were recruited through Amazon Mechanical Turk. Participants were paid $1.00 in compensation. In our preregistration, we indicated the following exclusion criteria: Participants were excluded from the study if they answered more than two out of eight manipulation check questions incorrectly. (These manipulation check questions directly queried whether the participants had read the stimuli text. Answering them incorrectly demonstrated lack of attention.) We determined the sample size via a power analysis based on the effect size of pilot studies and these exclusion criteria. In this power analysis, we calculated that we would need at least 27 participants to achieve 99% power with our pilot result effect size and α=0.01, using a Type-III *F*-test, but to ensure clear results chose to use a 480 as our sample size (one extra participant was recruited by the platform). However, our original exclusion criteria resulted in 212 participants being excluded (44% of our participants), which felt overly restrictive. For that reason, we decided to not use any exclusion criteria in Experiment 3, thereby including all participants in the analysis (we discuss the manipulation checks further in the Discussion). When we redid the power analysis calculation based on the pilot data with no exclusion criteria, using the same parameters as above, we calculated that we would need at least 179 participants to achieve 99% power. Thus, since 481 participants were included in the analysis, we still included a sample size with the desired power.

In the Supplemental Material, we restrict our sample to that specified by the preregistered exclusion criteria, and present the planned analyses and figures based on this sample for comparison. The results were similar to those presented here (Figure B.2).

Stimuli and Procedure

Video stimuli were the same as in Experiment 1.

After consenting, participants read introductory text: "In this experiment, you will see videos of someone choosing between an item in one bowl and an item in another bowl. There are 8 videos.

Please do your best to answer the questions afterward." Participants then entered the main experiment.

Participants either read descriptive text emphasizing that the decision-maker was feeling meticulous and cautious (high threshold condition), or that the decision-maker was feeling careless and inattentive (low threshold condition). The text for "high threshold" read: "This person is choosing between two items. It is very important to them to be very careful about the following decision. This person cares a lot about being *meticulous and cautious* right now. They had an easy day at work and so do not have much on their mind besides this decision." The text for "low threshold" read: "This person is choosing between two items. It is not important to them to be careful about the following decision. This person is fine being *careless and inattentive* right now. They had a hard day at work and have a lot on their mind besides this decision."[6]

Participants then saw a video underneath this text showing a decision-maker making a decision between two items, A and B, where the decision-maker took 3 seconds from video onset or 9 seconds from video onset to make these decisions. Next, participants were asked "How does the person feel right now?" and were given a binary choice between a "Meticulous and cautious" radio button and a "Careless and inattentive" radio button, which was used as a manipulation check to ensure the participants had read the descriptive text (the ordering of these buttons was random on each trial). Participants were finally shown the critical question, asking about the decision-maker's preferences: "How much do you think they value their chosen item?" Participants moved a slider containing values from 0 (labeled "Neutral between items") to 100 (labeled "Strongly prefers item"; 50 was labeled "Moderately prefers item"). No grid lines were shown, nor the numbers 0-100. The slider was initially set at 0, and participants were required to click and move the slider, but they could move it back to 0 if desired. After answering these two questions, participants could advance to the next page to see the next set of text/video/questions (Figure 3.8). There were eight pages total. All of the text/video/questions were on the same page, and participants could replay videos as often as they wished. Participants could not return to a previous page.

Each participant was randomly assigned to one of 12 counter-balancing variants in which the labeling and orientation of viewed videos were held fixed, and within which the participant viewed eight videos varying in response time and threshold. These eight videos presented in random order: Half of the choices were made in 3 seconds and half in 9 seconds (response time counterbalancing),

[6] This text was designed to isolate the threshold parameter as much as possible, and avoided e.g. mentioning differences in the absolute value of items, or response time cutoffs or pressure, for this reason. The information about having a hard day at work was included because the decision-maker appeared very focused in all video stimuli, and describing them as "careless" appeared unnatural otherwise.

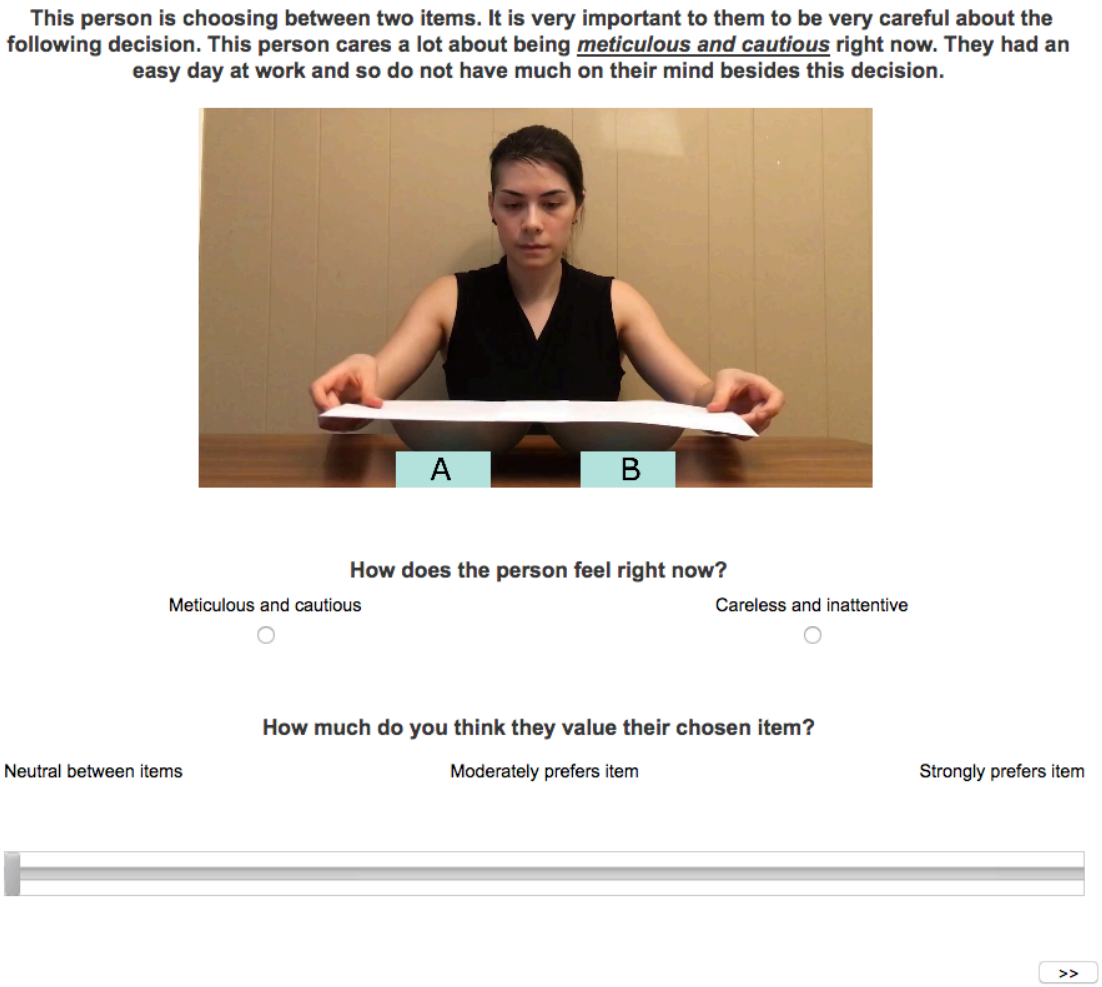


Figure 3.8: Experiment 3 Stimulus. Participants watched a 3- or 9-second video of the decision-maker choosing between A and B then answered questions.

half with the "high threshold" text and half with the "low threshold text" (threshold counterbalancing), and half with the decision-maker's right hand and half with the left hand (choice counterbalancing). After watching all eight videos and answering the accompanying questions, participants had the opportunity to write comments then exited the survey.

Model

We assume, as in Experiment 1, that participant responses are based on the posterior mean estimate of the difference in utilities of the two items. We additionally assume that the carefulness manipulation influences participants' assumed threshold, θ, such that those in the cautious condition will make predictions consistent with a high θ and those in the careless condition will make predictions consistent with a low θ. To capture this, we held fixed the β and α parameters fit in the previous two experiments, and fit θ separately to the two groups. We accounted for the different scale in the response slider (specific to Experiment 3) by setting the predicted response to $2 * \alpha\, \mathbb{E}[u_a - u_b \mid a \succ b, t]$. Besides these differences, the predictions were produced in the same way as for Experiment 1, using Equation 3.5.

3.6.2 Results

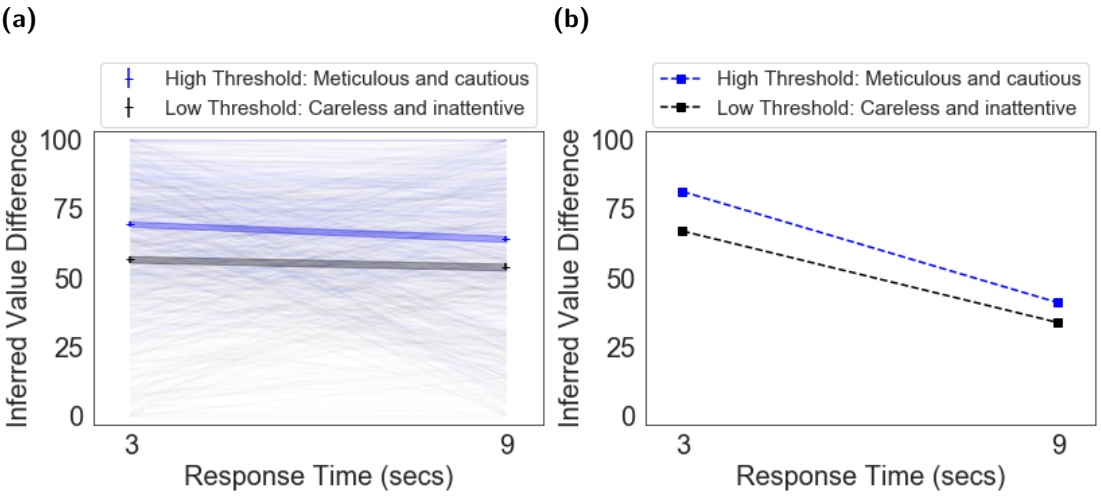


Figure 3.9: Experiment 3 Results. (a) Experimental Results. The mean ± SE of the inferred value difference, n=481, is shown for the high threshold ("cautious," blue) and the low threshold ("careless," black) conditions, for each of the 3 and 9 second response time conditions. Responses are connected by threshold condition for emphasis. Individual participants' averaged values for each threshold/time pair are also shown and connected. 0 represents the participant feeling that the decision-maker was neutral between items, 50 that the decision-maker moderately preferred their chosen item, and 100 that the decision-maker strongly preferred their chosen item. (b) Model Predictions.

In Experiment 3, we investigated whether participants could make inferences about decision-makers' preferences based on both timing and threshold (how careful the decision-maker is) simul-

taneously. Our model predicts a main effect of response time (the shorter the response time, the larger the inferred value difference between the two items will be) and a main effect of threshold (the higher the threshold, the larger the inferred value difference between the two items will be). Our model also predicts an interaction effect between response time and threshold, such that decision-makers who are cautious and fast to make decisions are predicted to particularly value their chosen item, compared to decision-makers who are cautious and slow to make decisions (whereas we expect a smaller difference for the response time variable if decision-makers are emphasized to be careless). Results are shown in Figure 3.9.

To analyze the results, we constructed a linear mixed effects model, with response time and threshold as categorical fixed effects, subject as a random effect, and participants' judgments of value (0-100, where 0 is "Neutral between items," 50 is "Moderately prefers item," and 100 is "Strongly prefers item") as the continuous dependent variable. The estimate (beta parameter) for the main effect of response time was 4.07 (95% CI=[2.62,5.53], t(3364.0)=5.48, p=4.5e-8, t-tests using Satterthwaite's method, η_p^2=0.009). The estimate for the main effect of threshold was 11.50 (95% CI=[10.04,12.95], t(3364.0)=15.48, p<2e-16, t-tests using Satterthwaite's method, η_p^2=0.07). The estimate for the interaction effect of response time and threshold was 2.62 (95% CI=[-0.29,5.53], t(3364.0)=1.76, p=0.08, t-tests using Satterthwaite's method, η_p^2=0.0009): a marginal but not significant effect. The estimated power of the interaction effect for α=0.01 was 19.50% (95% CI=[14.25,25.68]) using a Type-III F-test from the R package "car," generated via 200 simulations with the R package "simr." (The estimate for the fixed effect intercept was 61.13, 95% CI=[59.69,62.59], t(480.0)=82.97, p<2e-16, t-tests using Satterthwaite's method, η_p^2=0.94.) Results were similar, using the preregistered exclusion criteria; these results are included in the Supplemental Material (Figure B.2). Thus, while our empirical results matched our model predictions in that we observed the expected main effects of response time and threshold, we failed to observe the anticipated interaction effect.

3.6.3 Discussion

In Experiment 3, we asked whether participants could infer a decision-maker's preferences from their choices and response times while incorporating information about their mental state of being cautious or careless. As expected, we observed a main effect of response time in the expected direction (spending less time on a decision was associated with larger inferred preferences), though the main effect of response time was weaker than in Experiment 1, when there was only one manipulation (response time) rather than two (threshold and response time) and the critical question had a different format. We also observed a main effect of threshold in the expected direction (more care-

fulness was associated with larger inferred preferences). Finally, our model predicted an interaction effect, wherein a greater value difference was expected for cautious decision-makers with different response times compared to careless decision-makers, which we did not observe.

To manipulate threshold, we included a textual description above our video stimuli describing whether the decision-maker was feeling cautious or careless. However, in the videos, the decision-maker always appeared to be focused and therefore cautious, and participants expressed confusion in the experiment's comments section about the discrepancy between the textual description of the decision-maker's mental state and the decision-maker's facial expressions. This issue was reflected in the manipulation check questions, which asked "How does the person feel right now?" with the binary options of "Meticulous and cautious" and "Careless and inattentive." Participants asymmetrically answered the manipulation check questions incorrectly. When the described mental state was "Meticulous and cautious," 387/481 (80%) of participants answered at least 3 out of the 4 trials correctly (the mean correct number of trials was 3.3/4). When the described mental state was "Careless and inattentive," 261/481 (54%) of participants answered at least 3 out of the 4 trials correctly (the mean correct number of trials was 2.6/4).

In our preregistration, we had planned to exclude participants who answered more than two of the eight manipulation checks incorrectly, but these criteria would have resulted in 212 participants (44%) being excluded.[7] We thus included all participants in the experiment, though results were similar when the exclusion criteria were applied (Figure B.2). We expected that the high failure rate for the manipulation check exclusion criteria was due to the mismatch between the textual description of the decision-maker's mental state and the facial expression confound in the associated videos, so conducted Experiments 4A and 4B to address this confound.

3.7 Experiments 4A and 4B: Sensitivity to a decision-maker's mental state with different stimuli

In exploring whether knowing a decision-maker's mental state of cautious or careless influences people's inferences of their preferences, Experiment 3 introduced a confound in the threshold manipulation whereby the decision-maker's facial expressions did not match the textual descriptions of their mental state. We sought to address this using different stimuli in Experiment 4A (text message videos) and Experiment 4B (vignettes). In Experiment 4A, participants saw text message videos

[7] Note the locations of the manipulation check options were random for each trial, which likely contributed to some attention-based mistakes across both threshold conditions, separate from the described facial expression confound.

between the decision-maker and a friend, Alice. Alice either described the decision-maker as having a good day at work and careful in their decision-making (high threshold condition), or having a hard day at work (low threshold condition). Alice then asked the decision-maker to choose between items A and B. In the fast response condition, the decision-maker responded "Hm," then generated a decision. In the slow response condition, the decision-maker responded "Hm," then spent time deliberating, designated by a moving ellipsis, before giving a decision. In Experiment 4B, rather than watching a video, participants read a vignette of the decision-maker's mental state, the time they took to make a choice, and their choice. The manipulation check questions in Experiments 4A and 4B remained the same as those in Experiment 3. In both experiments, we addressed the same experimental question as in Experiment 3; thus, the model predictions are qualitatively the same.

3.7.1 Methods

Participants

In Experiment 4A, 480 participants were recruited through Prolific and paid $1.25 in compensation. In Experiment 4B, 481 participants were recruited through Prolific and were paid $1.00 in compensation. We did not exclude any participants. Experiments 4A and 4B used the same sample size and analyses as Experiment 3, and were not preregistered. Both experiments were completed at a later date and on a different platform (Prolific instead of Amazon Mechanical Turk), so it is possible but unlikely that participants from the first round of experiments (pilots, Experiments 1-3) participated in the second round of experiments (pilots, Experiments 4A and 4B). Participants within the second round of experiments could only complete one of that round's pilot experiments or main experiments.

Experiments 4A and 4B were IRB-approved by Princeton University, Protocol ID: 10859, Protocol Title: Computational Cognitive Science. Participants gave informed consent.

Stimuli and Procedure

The stimuli and procedure were similar to those in Experiment 3, but the main stimuli and critical question were changed. The main stimuli in Experiment 4A were changed to text message videos, and in Experiment 4B to vignettes. The critical question for Experiments 4A and 4B was posed as in Experiment 1 ("What are the person's feelings about item A and item B?") rather than Experiment 3 ("How much do you think they value their chosen item?"), since we considered the output of a scale spanning "Strongly Prefers A," "Neutral Between A and B," and "Strongly Prefers B" to be more

robust than Experiment 3's scale spanning "Neutral between items," "Moderately prefers [chosen] item," and "Strongly prefers [chosen] item."

After consenting, participants read introductory text. In Experiment 4A: "In this experiment, you will see text message videos of someone choosing between two items. There are 8 videos. Please do your best to answer the questions afterward." In Experiment 4B: "In this experiment, you will read vignettes of someone choosing between two items. There are 8 vignettes. Please do your best to answer the questions afterward." Participants then performed eight trials of the experimental task.

In Experiment 4A, on each trial participants read: "This person is choosing between two items. Please watch the video." A video underneath showed a text message conversation between Alice and the decision-maker. In the low threshold condition, the text in the video was the following. Alice: "I know you had a rough day, sorry to ask you this when you're distracted" / "but" / "of the two items we discussed" / "did you want" / "A or B?" The decision-maker then replied: "Hm" / "A" (or "B"). The slashes here represent line breaks. In the high threshold condition, the first two lines from Alice were replaced with: "I heard you had a nice day at work, so figured I'd ask you today since I know you're careful about this stuff :)". Pauses were placed before each new line to approximate typing and reading speed, and the total time of the videos was the same for the low and high threshold conditions. The response time manipulation was in how long the decision-maker took to answer "A" or "B." In the fast response condition, the answer came 2.0 seconds after "Hm" appeared. In the slow response condition, the answer came after 8.9 seconds; during this extended pause, an ellipsis (...) periodically displayed.

An example stimulus is shown in Figure 3.10.

In Experiment 4B, participants read: "Your friend had an easy day at work and enjoys being meticulous and cautious about their decisions. You've just asked them which of two items they'd prefer: A or B. After awhile, they answer A." This text describes the "high threshold" condition, and the slow response condition. In the "low threshold" condition, the first sentence was replaced with "Your friend had a hard day at work and is feeling careless and inattentive about their decisions." In the fast response condition, the last sentence was replaced with "Without pausing, they immediately answer A." Participants saw a total of eight vignettes: high/low threshold $\times$ slow/fast response $\times$ A/B choice. An example stimulus is shown in Figure 3.11.

Participants were then asked the manipulation check question: "How does the person feel right now?" Participants were given a binary choice between a "Meticulous and cautious" radio button and a "Careless and inattentive" radio button; the ordering of these buttons was randomized on each trial. Participants were finally shown the critical question, asking about the decision-maker's preferences: "What are the person's feelings about item A and item B?" Participants moved a slider

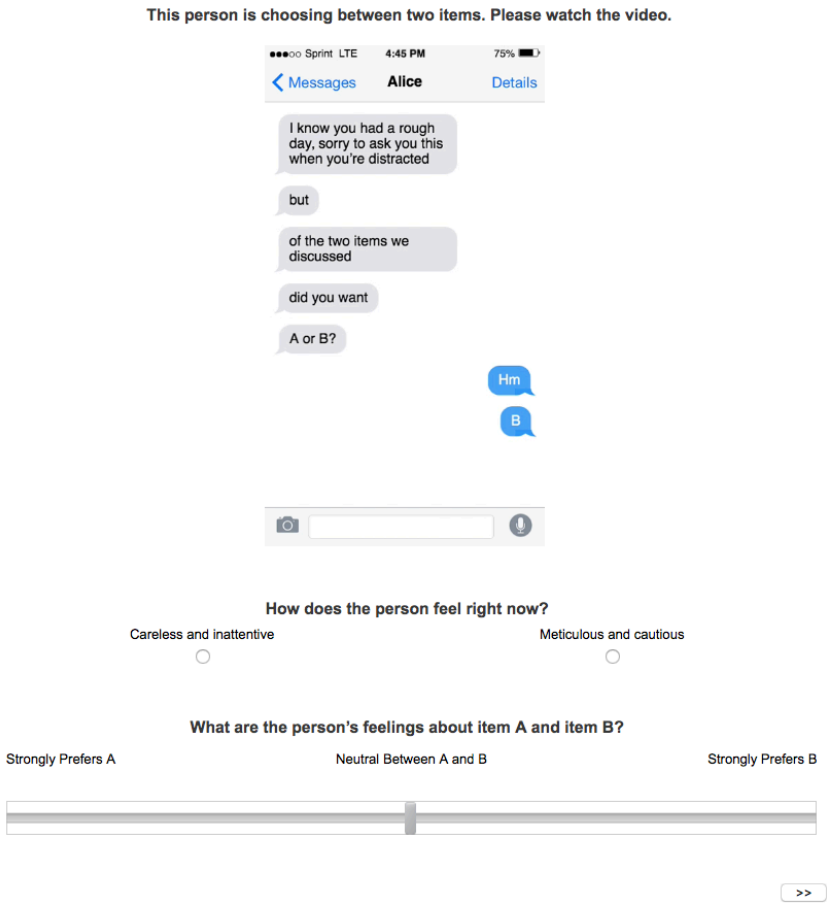


Figure 3.10: Experiment 4A Stimulus. Participants watched a text message video of a decision-maker choosing between A and B quickly or slowly, then answered questions. This image shows the final frame of the video.

containing values from 0 (labeled "Strongly Prefers A") to 100 (labeled "Strongly Prefers B"; 50 was labeled "Neutral Between A and B"). No grid lines were shown, nor the numbers 0-100. The slider was initially set at 50, and participants were required to click and move the slider, but they could move it back to 50 if desired. After answering these two questions, participants could advance to the next page. There were eight pages total. All of the text/(video)/questions were on the same page, and participants could replay videos as often as they wished if videos were present. Participants could not return to a previous page.

Each participant was randomly assigned to one of six counter-balancing variants in which the item labels were held fixed (AB, BA, CA, AC, BC, and CB). Only the AB variant, in which the

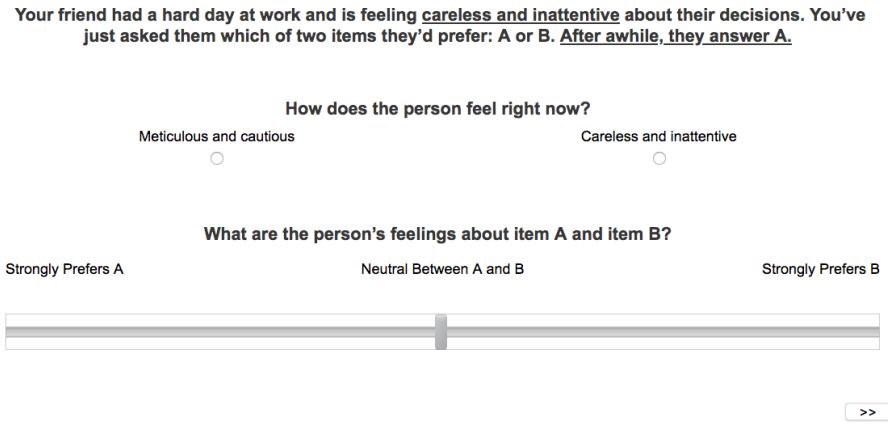


Figure 3.11: Experiment 4B Stimulus. Participants read vignettes of the decision-maker choosing between A and B quickly or slowly, then answered questions.

items were labeled A and B, will be described for expository purposes. Within each variant the participant viewed eight text message videos (Experiment 4A) or eight vignettes (Experiment 4B) varying in response time and threshold. These eight trials were presented in random order. Participants answered two questions for each of the combinations [cautious/fast], [careless/fast], [cautious/slow], [careless/slow]. After watching all eight text message videos or vignettes and answering the accompanying questions, participants had the opportunity to write comments then exited the survey.

3.7.2 Results

Here we asked whether people would incorporate a decision-maker's carefulness in inferring their preferences from response times, using different stimuli from Experiment 3: text message videos in Experiment 4A, and vignettes in Experiment 4B. The model predictions are identical to those for Experiment 3: we predicted a main effect of response time (the shorter the response time, the larger the inferred value difference between the two items) and a main effect of threshold (the higher the threshold, the larger the inferred value difference between the two items). We also predicted an interaction effect between response time and threshold, such that decision-makers who were cautious and fast to make decisions would be judged to particularly value their chosen item, compared to decision-makers who were cautious and slow to make decisions (whereas we would expect a smaller difference for the response time variable if decision-makers were emphasized to be careless). Results are shown in Figure 3.12.

To analyze the results, we created a linear mixed effects model, with response time and threshold

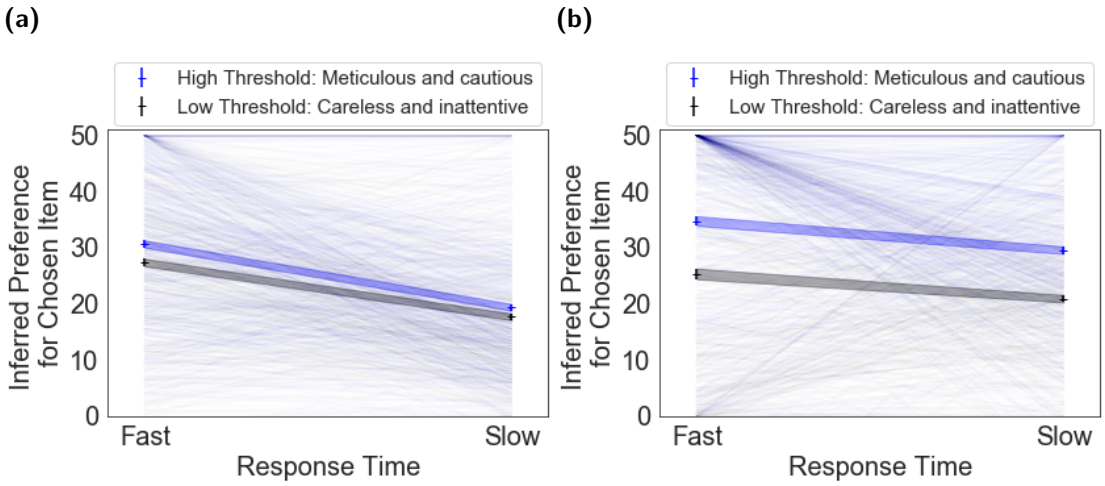


Figure 3.12: Experiment 4 Results. (a) Experiment 4A (Text Message Videos), n=480, and (b) Experiment 4B (Vignettes), n=481. The mean ± SE of inferred value difference is shown for the high threshold ("cautious," blue) and the low threshold ("careless," black) conditions, for each of the fast and slow response time conditions. Responses were reoriented to be 0-50: 0 indicates the participant felt the decision-maker was neutral between items A and B, and 50 indicates the participant felt the decision-maker strongly preferred the chosen item (A or B, whichever was appropriate to the trial). Responses are connected by threshold condition for emphasis. Individual participants' averaged values for each threshold/time pair are also shown and connected.

as categorical fixed effects, subject as a random effect, and participants' preference judgments (0-100, 50 was "Neutral Between A and B," 0 was "Strongly Prefers A" and 100 was "Strongly Prefers B") as the continuous output variable.

For Experiment 4A, the results from the linear mixed effects model were as follows. The estimate (beta parameter) for the main effect of response time was 10.49 (95% CI=[9.74,11.24], t(3357.0)=27.36, p<2e-16, t-tests using Satterthwaite's method, η_p^2=0.18). The estimate for the main effect of threshold was 2.50 (95% CI=[1.75,3.26], t(3357.0)=6.53, p=8e-11, t-tests using Satterthwaite's method, η_p^2=0.01). The estimate for the interaction effect of response time and threshold was 1.59 (95% CI=[0.086,3.09], t(3357.0)=2.07, p=0.038, t-tests using Satterthwaite's method, η_p^2=0.001): a significant effect. The estimated power of the interaction effect for α=0.01 was 30.00% (95% CI=[23.74,36.86]) using a Type-III F-test from the R package "car," generated via 200 simulations with the R package "simr." (The estimate for the fixed effect intercept was 23.78, 95% CI=[22.84,24.73], t(479.0)=49.58, p<2e-16, t-tests using Satterthwaite's method, η_p^2=0.84.) Thus, our empirical results matched our model predictions in that we observed the expected main effects of response time and threshold. We also observed the predicted interaction effect (though

only at 30.00% estimated power).

Participants were also asked a manipulation check question: "How does the person feel right now?" Participants had to choose between "Meticulous and cautious" and "Careless and inattentive." (The locations of these options was random on every trial, which likely contributed to some mistakes.) However, participants asymmetrically answered the manipulation check questions incorrectly in Experiment 4A. When the described mental state was "Meticulous and cautious," 321/480 (67%) of participants answered at least 3 out of the 4 trials correctly (the mean correct number of trials was 2.9/4). When the described mental state was "Careless and inattentive," 121/480 (25%) of participants answered at least 3 out of the 4 trials correctly (the mean correct number of trials was 1.8/4). Under the preregistered exclusion criteria specified for Experiment 3, 341 participants (71%) would have been excluded for answering more than two of the eight manipulation checks incorrectly. Thus, we elected to not exclude participants based on their performance on the manipulation check (nor for any other reason).

For Experiment 4B, the results from the linear mixed effects model were as follows. The estimate (beta parameter) for the main effect of response time was 4.81 (95% CI=[3.71,5.91], t(3364.0)=8.54, p<2e-16, t-tests using Satterthwaite's method, η_p^2=0.02). The estimate for the main effect of threshold was 9.03 (95% CI=[7.93,10.14], t(3364.0)=16.06, p<2e-16, t-tests using Satterthwaite's method, η_p^2=0.07). The estimate for the interaction effect of response time and threshold was 0.79 (95% CI=[-1.42,2.99], t(3364.0)=0.70, p=0.48, t-tests using Satterthwaite's method, η_p^2=0.0001): this effect was not present. The estimated power of the interaction effect for α=0.01 was 2.50% (95% CI=[0.82,5.74]) using a Type-III F-test from the R package "car," generated via 200 simulations with the R package "simr." (The estimate for the fixed effect intercept was 27.66, 95% CI=[26.67,28.64], t(480.0)=54.94, p<2e-16, t-tests using Satterthwaite's method, η_p^2=0.86.) Thus, while we observed the expected main effects of response time and threshold, the predicted interaction effect was not present in Experiment 4B.

Participants did not answer the manipulation check questions asymmetrically in Experiment 4B. When the described mental state was "Meticulous and cautious," 369/481 (77%) of participants answered at least 3 out of the 4 trials correctly (the mean correct number of trials was 3.38/4). When the described mental state was "Careless and inattentive," 379/481 (79%) of participants answered at least 3 out of the 4 trials correctly (the mean correct number of trials was 3.44/4). Under the preregistered exclusion criteria specified for Experiment 3, 101 participants (21%) would have been excluded for answering more than two of the eight manipulation checks incorrectly. We did not exclude these participants, however, for consistency with Experiment 4A.

3.7.3 Discussion

The pattern of results across Experiments 3, 4A, and 4B support our model's prediction that people can infer decision-makers' preferences based on response times in a way that is appropriately sensitive to the decision-maker's mental state. First, we observed a main effect of response time across all three experiments, even though the response time manipulation was implemented very differently for each (time for a decision-maker to visually reach for a bowl in Experiment 3, time to type a text message response in Experiment 4A, and "After awhile, they answer A," or "Without pausing, they immediately answer A," in Experiment 4B). The main effect of response time was strongest in Experiment 4A, followed by Experiment 4B, and weakest in Experiment 3. One difference between Experiment 3 and Experiments 4A and 4B is that Experiment 3 had a different critical question, but even if it had had the same critical question, Experiments 4A and 4B had distinct enough patterns of results from each other (response time, threshold, interaction effects) that it is hard to know how this switch would have affected the results. Regardless, we observed a significant effect of response time across Experiments 1, 2, 3, 4A, and 4B, whereby participants inferred that a decision-maker had a stronger preference for an item the less time the decision-maker spent choosing it, supporting the robustness of the effect.

We also observed a main effect of threshold across Experiments 3, 4A, and 4B, in that if the decision-maker was described as cautious, participants inferred that the decision-maker had a stronger preference for their chosen item. This threshold effect was less present in Experiment 4A compared to Experiments 4B and 3, likely because the threshold manipulation was less explicitly emphasized (participants had to infer the decision-maker's mental state from Alice's description). However, we find it reassuring that a threshold manipulation as minor (and naturalistic) as that in Experiment 4A was enough to influence people's inferences about the decision-maker's preferences. We believe that the presence of both main effects—threshold and response time—across quite different paradigms speaks to the strength of people's ability to socially infer preferences from response times while taking into account decision-makers' mental states.

The model also predicted an interaction effect, wherein the effect of response time on inferred value is larger for cautious vs. careless decision-makers. This interaction effect did not occur at all in Experiment 4B (p=0.48, 2.5% power), was only marginally significant in Experiment 3 (p=0.08, 19.5% power), and was significant at low power in Experiment 4A (p=0.04, 30% power). Further work will be required to uncover how people jointly reason about how careful a decision-maker is and their response times in inferring the decision-maker's preferences.

Finally, manipulation check performance was strongest and symmetric across threshold condi-

tions in Experiment 4B, worse and asymmetric in Experiment 3, and poorest and asymmetric in Experiment 4A. Experiments 4B and 3 contained explicit threshold manipulations so it is not surprising that their manipulation check performance was higher than in Experiment 4A, in which the threshold manipulation was weaker. It was also expected that Experiment 4B would have higher and more symmetric manipulation check performance compared to Experiment 3, because Experiment 4B did not contain a facial expression confound. It is surprising, however, that Experiment 4A had asymmetric performance across the threshold conditions because the stimuli were texts without a facial expression confound. One explanation is that the high threshold condition in Experiment 4A contained the word "careful" (which is semantically close to the high-threshold manipulation check option "Meticulous and cautious"), while the low threshold condition only included the word "distracted" (which is arguably less close to the low-threshold manipulation check option "Careless and inattentive"). Alternatively, perhaps our stimuli generally resulted in a bias for attributing carefulness to the decision-maker; Convincing participants that the decision-maker was being at least somewhat deliberate in their choices was a requisite property in our experiments. To conclude, the fact that we observed main effects of response time and threshold across all three experiments suggests that manipulation check performance is informative but not central to our model's predictions and results.

3.8 General Discussion

If your friend takes a long time to decide between two options, you can infer that both options have similar value to her. Making these kinds of inferences about others' preferences is ubiquitous in daily life as people seamlessly integrate choices and response times. In this work, we created a computational model to describe this phenomenon by inverting a classic generative model of decision-making, the Drift Diffusion Model. We tested the predictions of this model for three experimental questions.

In Experiment 1, participants inferred a decision-maker preferred a chosen item more when they spent a longer time deciding between two items, matching model predictions. In Experiment 2, participants inferred a decision-maker's relative preferences from two choices and response times to infer a decision-maker's preference about a third unseen choice, and they inferred the model's predicted choice greater than chance across all conditions. Experiment 2 also included two control conditions in which participants only needed choice information to anticipate the decision-maker's unseen choice. As predicted, participants thought the predicted choice was more likely in these control conditions, providing an upper bound for participant likelihoods in this experiment. On

the other hand, participant behavior for the different conditions was more variable than expected, which should be incorporated into the model in future work. In Experiment 3, we asked whether participants could integrate the decision-maker's mental state of cautiousness or carelessness when inferring preferences. We observed the expected main effects of threshold and response time in three experiments spanning a range of stimuli (Experiments 3, 4A, and 4B), but a more complicated pattern of results in observing the expected interaction effect. Future work should investigate how experimental manipulations of threshold and response times are entangled in inferences about decision-makers' preferences.

The uniqueness of this work lies in the question asked and the inversion of the DDM to formally characterize our predictions. Specifically, we asked how people infer other's preferences by observing their response times and choices, which involves characterizing people's theory of mind surrounding response time. The work closest to this appears in a recent preprint by Konovalov and Krajbich (2020), who demonstrate that people use response times to make inferences about others' preferences, and further than people apply this knowledge strategically in bargaining games. Where our work differs is we formally describe these inferences about other's minds by inverting the DDM and making direct, quantitative predictions, whereas Konovalov and Krajbich (2020) make qualitative predictions based on the DDM. (Konovalov and Krajbich (2020) also apply their work to strategic settings, while we explore prediction for novel choices and integrating the decision-maker's mental state of cautiousness or carelessness.)

Our work has several potential extensions. Previous work already touches on interesting applications in strategic gameplay and mechanism design: Frydman and Krajbich (2018) designed an experiment in which some participants were explicitly shown others' response times for decision-making, and showed that those participants used their response time inferences about others' preferences to achieve better task outcomes compared to participants who did not have access to response time information. Expanding upon this, Konovalov and Krajbich (2020) showed that participants improved their outcomes by integrating either explicitly-presented or real-time perception of response time in a bargaining game. There is a fascinating aspect of mechanism design in these studies, in that people's performance is improved by the researchers introducing additional information—response times—from which players make inferences. This aspect of mechanism design is even more prominent in Krajbich et al. (2014),[8] albeit Krajbich et al. (2014) did not involve participants inferring other people's preferences. Krajbich et al. (2014) expected based on DDM predictions that participants would spend more time on difficult choices—in which options were of similar value—even when opportunity costs suggested participants should pick one and move on. The authors designed

[8] See also Oud et al. (2016).

an experiment where opportunity cost was high, and then implemented an intervention to end trials early when participants took too long to respond (an intervention they anticipated would target difficult choices in which participants were close to indifference). As predicted, after experiencing this intervention participants went on to earn higher rewards on the task on non-intervention blocks, from which the authors conclude that "it may be possible to improve people's welfare with simple interventions or behavioral training." In our work, we created an inverse DDM that generates predictions about how people infer other's preferences from response times. One could imagine that this output could be a valuable tool in mechanism design: Presenting the output from such a model in strategic games, either as a training intervention or throughout the whole task, could serve as an assistive tool to improve people's performance. Integrating mechanism design for inference of others' preferences from response times in real world settings would be an intriguing application for this work.

The inverted DDM captures some elements of social inference from response times, but a useful future direction would be to incorporate the idea of the overall value of choice sets, and other elements of "context." For example, the DDM describes a decision-maker's difference in value between two items, but it does not incorporate the idea that choice sets can be of overall high value (e.g. buying a house, going to college) or low value (e.g. what candy bar to eat). Empirically, as the overall value of a choice set increases, response times decrease; fortunately, researchers have developed alternative sequential sampling models that can account for this effect (Clithero, 2018b; Shevlin & Krajbich, 2020). Incorporating "context" factors in inverted DDMs such as overall value, choice option similarity (Bhatia & Mullett, 2018), attention (Krajbich, 2019) and (as investigated in this work) mental states can not only help us describe and predict people's behavior, but could be used as tools to help nudge people towards more optimal use of their time.

Finally, the particular inversion of the DDM in this work has applications for human-robot interaction. This work could be interpreted as inverse reinforcement learning (Jara-Ettinger, 2019; Ng & Russell, 2000) with algorithm run times: inferring someone's utility function based on how long it takes them to decide. Artificial decision-makers are generally quite poor at social inference, especially compared to human performance, and so incorporating formal models how people learn about others through response times will only improve communication between people and automated systems.

3.9 Conclusion

If your friend immediately says "Yes!" to a movie but contemplates before saying "...yes" to hiking, you've learned something about the relative strength of her preferences that you couldn't learn just from her choices. People constantly use response time to make inferences about each other's preferences in daily life, and in this work we sought to quantitatively capture and predict that phenomenon. We present a rational model of inferring preferences from response time, using the Drift Diffusion Model to describe how preferences influence response time and Bayesian inference to invert this relationship. We use our inverted DDM to predict participant behavior for three experimental questions. We first demonstrated that people could infer that a decision-maker preferred a chosen item more if they spent longer deliberating (Experiment 1). We then showed that people could predict 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). Finally we observed that people could incorporate information on a decision-maker's mental state, cautiousness or carelessness, in inferring a decision-maker's preferences from their response times (Experiments 3, 4A, and 4B).

We live in a social world, and people perform a startling number of unconscious inferences in navigating others' preferences and goals. While economists and psychologists have been characterizing how we learn about others' minds from their choices for decades, capturing the more subtle inferences made from watching someone think is also essential to understanding how we are so good at knowing what is not said. In this paper we advance this goal by creating a rational model of inverted decision-making to describe these inferences from response times. Much remains to be incorporated into these models. Within our results, more remains to be investigated around how preferences are inferred if decision-makers' mental states are presented in ways that interact with cues like response time. Within DDMs, response times are influenced by other factors than just value difference, such as overall value of choice sets and many different types of "context" (we investigated one version of context, a decision-maker's mental state of cautiousness or carelessness). Outside of DDMs, incorporating tone in addition to choices and response times will be key to describing how people make inferences about others' preferences. Fortunately, the applications for formalizing models of social inferences are exciting and broad. People already draw impressive conclusions about others' beliefs, preferences, and future actions from their choices and response times, and fully elucidating models of these social inferences not only helps us understand the mechanisms behind such powerful computational process, but may help us help people make better choices—via mechanism design, or building tools based on humans' existing capacities—in the future.

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 them-

selves, 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, Sevenhant, & Roets, 2018; FeldmanHall, Dalgleish, et al., 2012; FeldmanHall, Mobbs, et al., 2012; Tassy, Oullier, Mancini, & Wicker, 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, van den Bos, Rombouts, and Crone (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, Harre, and Lopez (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, Heidari, and Krause (2019) and Yaghini, Heidari, and Krause (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, Zaki, & Goodman, 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, Goodman, & Griffiths, 2014), evaluating deception (Shafto, Eaves, Navarro, & Perfors, 2012; Sobel & Kushnir, 2013), pragmatics (Goodman & Frank, 2016), and many other areas (Jara-Ettinger, 2019; Shum, Kleiman-

Weiner, Littman, & Tenenbaum, 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

Acuna, B. D., Sanes, J. N., & Donoghue, J. P. (2002). Cognitive mechanisms of transitive inference. *Experimental Brain Research*, *146*(1), 1–10. doi:10.1007/s00221-002-1092-y

Ahlert, M., Funke, K., & Schwettmann, L. (2013). Thresholds, productivity, and context: An experimental study on determinants of distributive behaviour. *Social Choice and Welfare*, *40*(4), 957–984. doi:https://doi.org/10.1007/s00355-012-0652-8

Aleksandrov, M. D., Aziz, H., Gaspers, S., & Walsh, T. (2015). Online fair division: Analysing a food bank problem. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*.

Alesina, A., & Angeletos, G.-M. (2005). Fairness and redistribution. *American Economic Review*, *95*(4), 960–980. doi:https://doi.org/10.1257/0002828054825655

Alós-Ferrer, C., & Garagnani, M. (2020). Strength of preference and decision making under risk. *University of Zurich, Department of Economics, Working Paper*, (330). doi:10.2139/ssrn.3428515

Amanatidis, G., Markakis, E., Nikzad, A., & Saberi, A. (2017). Approximation algorithms for computing maximin share allocations. *ACM Transactions on Algorithms*, *13*(4), 52. doi:https://doi.org/10.1145/3147173

Amasino, D. R., Sullivan, N. J., Kranton, R. E., & Huettel, S. A. (2019). Amount and time exert independent influences on intertemporal choice. *Nature Human Behaviour*, *3*(4), 383–392. doi:10.1038/s41562-019-0537-2

Andersson, F., & Lyttkens, C. H. (1999). Preferences for equity in health behind a veil of ignorance. *Health Economics*, *8*(5), 369–378. doi:https://doi.org/10.1002/(SICI)1099-1050(199908)8:5<369::AID-HEC456>3.0.CO;2-Q

Andreoni, J. (1989). Giving with impure altruism: Applications to charity and Ricardian equivalence. *Journal of Political Economy*, *97*(6), 1447–1458. doi:https://doi.org/10.1086/261662

Andreoni, J., Miller, J. H. et al. (1993). Rational cooperation in the finitely repeated prisoner's dilemma: Experimental evidence. *Economic Journal*, *103*(418), 570–85. doi:https://doi.org/10.2307/2234532

Andreoni, J., & Vesterlund, L. (2001). Which is the fair sex? Gender differences in altruism. *The Quarterly Journal of Economics*, *116*(1), 293–312. doi:https://doi.org/10.1162/003355301556419

Ashlagi, I., Karagözoğlu, E., & Klaus, B. (2012). A non-cooperative support for equal division in estate division problems. *Mathematical Social Sciences*, *63*(3), 228–233. doi:https://doi.org/10.1016/j.mathsocsci.2012.01.004

Austerweil, J. L., Brawner, S., Greenwald, A., Hilliard, E., Ho, M., Littman, M. L., ... Trimbach, C. (2015). The impact of other-regarding preferences in a collection of non-zero-sum grid games. *AAAI Spring Symposium 2016 on Challenges and Opportunities in Multiagent Learning for the Real World.* Palo Alto, CA: The AAAI Press.

Babcock, L., Loewenstein, G., Issacharoff, S., & Camerer, C. (1995). Biased judgments of fairness in bargaining. *American Economic Review*, *85*(5), 1337–1343.

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. doi:10.1038/s41562-017-0064

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

Barman, S., & Krishna Murthy, S. K. (2017). Approximation algorithms for maximin fair division. In *Proceedings of the 2017 ACM Conference on Economics and Computation* (pp. 647–664). ACM. doi:https://doi.org/10.1145/3033274.3085136

Basden, B., Basden, D., Bryner, S., & Thomas, R. (1997). A comparison of group and individual remembering: Does collaboration disrupt retrieval strategies? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*(5), 1176.

Basden, B., Basden, D., & Henry, S. (2000). Costs and benefits of collaborative remembering. *Applied Cognitive Psychology*, *14*(6), 497–507.

Beckman, S. R., Formby, J. P., Smith, W. J., & Zheng, B. (2002). Envy, malice and Pareto efficiency: An experimental examination. *Social Choice and Welfare*, *19*(2), 349–367. doi:https://doi.org/10.1007/s003550100116

Berg, J., Dickhaut, J., & McCabe, K. (1995). Trust, reciprocity, and social history. *Games and Economic Behavior*, *10*(1), 122–142. doi:https://doi.org/10.1006/game.1995.1027

Bergemann, D., & Valimaki, J. (2006). Efficient dynamic auctions. doi:https://doi.org/10.2139/ssrn.936633

Bertsimas, D., Farias, V. F., & Trichakis, N. (2012). On the efficiency-fairness trade-off. *Management Science*, *58*(12), 2234–2250. doi:https://doi.org/10.1287/mnsc.1120.1549

Bertsimas, D., Farias, V., & Trichakis, N. (2011). The price of fairness. *Operations Research*, *59*(1), 17–31. doi:https://doi.org/10.1287/opre.1100.0865

Bhatia, S., & Mullett, T. L. (2018). Similarity and decision time in preferential choice. *Quarterly Journal of Experimental Psychology*, *71*(6), 1276–1280. doi:10.1177/1747021818763054

Binmore, K. (1994). *Game Theory and the Social Contract. Vol 1, Playing Fair (MIT Press Series on Economic Learning and Social Evolution)*. MIT Press.

Bitzer, S., Park, H., Blankenburg, F., & Kiebel, S. J. (2014). Perceptual decision making: drift-diffusion model is equivalent to a Bayesian model. *Frontiers in Human Neuroscience*, *8*, 102. doi:10.3389/fnhum.2014.00102

Bogacz, R., Brown, E., Moehlis, J., Holmes, P., & Cohen, J. D. (2006). The physics of optimal decision making: A formal analysis of models of performance in two-alternative forced-choice tasks. *Psychological Review*, *113*(4), 700. doi:10.1037/0033-295X.113.4.700

Bolton, G., Brandts, J., Katok, E., Ockenfels, A., & Zwick, R. (2008). Testing theories of other-regarding behavior: A sequence of four laboratory studies. *Handbook of Experimental Economics Results*, *1*, 488–499. doi:https://doi.org/10.1016/S1574-0722(07)00055-8

Bolton, G., & Ockenfels, A. (2000). ERC: A theory of equity, reciprocity, and competition. *American Economic Review*, *90*(1), 166–193. doi:https://doi.org/10.1257/aer.90.1.166

Bosmans, K., & Schokkaert, E. (2004). Social welfare, the veil of ignorance and purely individual risk: An empirical examination. In *Inequality, Welfare and Income Distribution: Experimental Approaches* (pp. 85–114). Emerald Group Publishing Limited.

Bosmans, K., & Schokkaert, E. (2009). Equality preference in the claims problem: A questionnaire study of cuts in earnings and pensions. *Social Choice and Welfare*, *33*(4), 533. doi:https://doi.org/10.1007/s00355-009-0378-4

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.

Bouveret, S., & Lang, J. (2011). A general elicitation-free protocol for allocating indivisible goods. In *Twenty-Second International Joint Conference on Artificial Intelligence*.

Boyd, R., & Lorberbaum, J. P. (1987). No pure strategy is evolutionarily stable in the repeated prisoner's dilemma game. *Nature*, *327*(6117), 58. doi:https://doi.org/10.1038/327058a0

Brams, S. J., Edelman, P. H., & Fishburn, P. C. (2003). Fair division of indivisible items. *Theory and Decision*, *55*(2), 147–180. doi:https://doi.org/10.1023/B:THEO.0000024421.85722.0a

Brams, S. J., & Taylor, A. D. (1996). *Fair Division: From cake-cutting to dispute resolution*. Cambridge University Press.

Brunton, B. W., Botvinick, M. M., & Brody, C. D. (2013). Rats and humans can optimally accumulate evidence for decision-making. *Science*, *340*(6128), 95–98. doi:10.1126/science.1231965

Busemeyer, J. R. (1985). Decision making under uncertainty: A comparison of simple scalability, fixed-sample, and sequential-sampling models. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *11*(3), 538. doi:10.1037/0278-7393.11.3.538

Busemeyer, J. R., & Rieskamp, J. (2014). Psychological research and theories on preferential choice. In *Handbook of choice modelling*. doi:10.4337/9781781003152.00008

Busemeyer, J. R., & Townsend, J. T. (1993). Decision field theory: A dynamic-cognitive approach to decision making in an uncertain environment. *Psychological Review*, *100*(3), 432. doi:10.1037//0033-295X.100.3.432

Camerer, C. F. (2011). *Behavioral game theory: Experiments in strategic interaction*. Princeton University Press.

Cappelen, A. W., Hole, A. D., Sørensen, E. Ø., & Tungodden, B. (2007). The pluralism of fairness ideals: An experimental approach. *American Economic Review*, *97*(3), 818–827. doi:https://doi.org/10.1257/aer.97.3.818

Cappelen, A. W., Nielsen, U. H., Sørensen, E. Ø., Tungodden, B., & Tyran, J.-R. (2013). Give and take in dictator games. *Economics Letters*, *118*(2), 280–283. doi:https://doi.org/10.1016/j.econlet.2012.10.030

Carlsson, F., Gupta, G., & Johansson-Stenman, O. (2003). Choosing from behind a veil of ignorance in India. *Applied Economics Letters*, *10*(13), 825–827. doi:https://doi.org/10.1080/1350485032000148268

Cavallo, R. (2008). Efficiency and redistribution in dynamic mechanism design. In *Proceedings of the Ninth ACM Conference on Electronic Commerce* (pp. 220–229). ACM. doi:https://doi.org/10.1145/1386790.1386826

Chabris, C. F., Laibson, D., Morris, C. L., Schuldt, J. P., & Taubinsky, D. (2008). *Measuring intertemporal preferences using response times*. National Bureau of Economic Research. doi:10.3386/w14353

Chabris, C. F., Morris, C. L., Taubinsky, D., Laibson, D., & Schuldt, J. P. (2009). The allocation of time in decision-making. *Journal of the European Economic Association*, *7*(2-3), 628–637. doi:10.1162/jeea.2009.7.2-3.628

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

Charness, G., & Rabin, M. (2002). Understanding social preferences with simple tests. *The Quarterly Journal of Economics*, *117*(3), 817–869. doi:https://doi.org/10.1162/003355302760193904

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

Chmura, T., Kube, S., Pitz, T., & Puppe, C. (2005). Testing (beliefs about) social preferences: Evidence from an experimental coordination game. *Economics Letters*, *88*(2), 214–220. doi:https://doi.org/10.1016/j.econlet.2005.02.009

Clithero, J. A. (2018a). Improving out-of-sample predictions using response times and a model of the decision process. *Journal of Economic Behavior & Organization*, *148*, 344–375. doi:10.1016/j.jebo.2018.02.007

Clithero, J. A. (2018b). Response times in economics: Looking through the lens of sequential sampling models. *Journal of Economic Psychology*, *69*, 61–86. doi:10.1016/j.joep.2018.09.008

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.

Cooney, G., Gilbert, D., & Wilson, T. (2016). When fairness matters less than we expect. *Proceedings of the National Academy of Sciences*, *113*(40), 11168–11171. doi:https://doi.org/10.1073/pnas.1606574113

Cox, J. C. (2004). How to identify trust and reciprocity. *Games and Economic Behavior*, *46*(2), 260–281. doi:https://doi.org/10.1016/S0899-8256(03)00119-2

Croson, R., & Gneezy, U. (2009). Gender differences in preferences. *Journal of Economic Literature*, *47*(2), 448–74. doi:https://doi.org/10.1257/jel.47.2.448

Croson, R., & Konow, J. (2009). Social preferences and moral biases. *Journal of Economic Behavior & Organization*, *69*(3), 201–212.

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

Dai, J., & Busemeyer, J. R. (2014). A probabilistic, dynamic, and attribute-wise model of intertemporal choice. *Journal of Experimental Psychology: General*, *143*(4), 1489. doi:http://dx.doi.org/10.3758

Dana, J., Weber, R. A., & Kuang, J. X. (2007). Exploiting moral wiggle room: Experiments demonstrating an illusory preference for fairness. *Economic Theory*, *33*(1), 67–80. doi:https://doi.org/10.1007/s00199-006-0153-z

Demers, A., Keshav, S., & Shenker, S. (1989). Analysis and simulation of a fair queueing algorithm. In *ACM SIGCOMM Computer Communication Review* (Vol. 19, *4*, pp. 1–12). ACM.

Deutsch, M. (1985). Distributive justice: A social-psychological perspective.

Dickerson, J. P., Goldman, J. R., Karp, J., Procaccia, A. D., & Sandholm, T. (2014). The computational rise and fall of fairness. In *AAAI* (Vol. 14, pp. 1405–1411).

Diederich, A. (1997). Dynamic stochastic models for decision making under time constraints. *Journal of Mathematical Psychology*, *41*(3), 260–274. doi:10.1006/jmps.1997.1167

Diederich, A. (2003). Decision making under conflict: Decision time as a measure of conflict strength. *Psychonomic Bulletin & Review*, *10*(1), 167–176. doi:10.3758/BF03196481

Dreber, A., Fudenberg, D., & Rand, D. G. (2014). Who cooperates in repeated games: The role of altruism, inequity aversion, and demographics. *Journal of Economic Behavior & Organization*, *98*, 41–55. doi:https://doi.org/10.1016/j.jebo.2013.12.007

Dubois, D., Fargier, H., & Prade, H. (1996). Refinements of the maximin approach to decision-making in a fuzzy environment. *Fuzzy Sets and Systems*, *81*(1), 103–122. doi:https://doi.org/10.1016/0165-0114(95)00243-X

Dufwenberg, M., & Kirchsteiger, G. (2004). A theory of sequential reciprocity. *Games and Economic Behavior*, *47*(2), 268–298. doi:https://doi.org/10.1016/j.geb.2003.06.003

Dupuis-Roy, N., & Gosselin, F. (2009). An empirical evaluation of fair-division algorithms. In *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (pp. 2681–2686).

Dupuis-Roy, N., & Gosselin, F. (2011). The simpler, the better: A new challenge for fair-division theory. In *Proceedings of the 33rd Annual Meeting of the Cognitive Science Society* (pp. 3229–3234).

Echenique, F., & Saito, K. (2017). Response time and utility. *Journal of Economic Behavior & Organization*, *139*, 49–59. doi:10.1016/j.jebo.2017.04.008

Ellingsen, T., & Johannesson, M. (2001). Sunk Costs, Fairness, and Disagreement. *Stockholm School of Economics Working Paper*.

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

Escoffier, B., Gourvès, L., & Monnot, J. (2013). Fair solutions for some multiagent optimization problems. *Autonomous Agents and Multi-Agent Systems*, *26*(2), 184–201. doi:https://doi.org/10.1007/s10458-011-9188-z

Faravelli, M. (2007). How context matters: A survey based experiment on distributive justice. *Journal of Public Economics*, *91*(7-8), 1399–1422. doi:https://doi.org/10.1016/j.jpubeco.2007.01.004

Fehr, E. [E.], Naef, M., & Schmidt, K. (2006). Inequality aversion, efficiency, and maximin preferences in simple distribution experiments: Comment. *American Economic Review*, *96*(5), 1912–1917. doi:https://doi.org/10.1257/aer.96.5.1912

Fehr, E. [Ernst], & 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.

Finlay, F., Hitch, G., & Meudell, P. (2000). Mutual inhibition in collaborative recall: evidence for a retrieval-based account. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(6), 1556.

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

Fisman, R., Kariv, S., & Markovits, D. (2007). Individual preferences for giving. *American Economic Review*, *97*(5), 1858–1876. doi:https://doi.org/10.1257/aer.97.5.1858

Fleurbaey, M. (2008). *Fairness, responsibility, and welfare*. doi:https://doi.org/10.1093/acprof:osobl/9780199215911.001.0001

Fong, C. (2001). Social preferences, self-interest, and the demand for redistribution. *Journal of Public Economics*, *82*(2), 225–246. doi:https://doi.org/10.1016/S0047-2727(00)00141-9

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

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

Freeman, R., Zahedi, S. M., & Conitzer, V. (2017). Fair social choice in dynamic settings. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI). Forthcoming*. doi:https://doi.org/10.24963/ijcai.2017/639

Frohlich, N., & Oppenheimer, J. A. (1992). Choosing justice: An experimental approach to ethical theory. University of California Press.

Frohlich, N., Oppenheimer, J. A., & Eavey, C. L. (1987). Laboratory results on Rawls's distributive justice. *British Journal of Political Science*, *17*(1), 1–21. doi:https://doi.org/10.1017/S0007123400004580

Frydman, C., & Krajbich, I. (2018). Using response times to infer others' beliefs: An application to information cascades. *Available at SSRN 2817026*. doi:10.2139/ssrn.2817026

Gächter, S., & Riedl, A. (2006). Dividing justly in bargaining problems with claims. *Social Choice and Welfare*, *27*(3), 571–594. doi:https://doi.org/10.1007/s00355-006-0141-z

Gaertner, W., Jungeilges, J., & Neck, R. (2001). Cross-cultural equity evaluations: A questionnaire-experimental approach. *European Economic Review*, *45*(4-6), 953–963. doi:https://doi.org/10.1016/S0014-2921(01)00119-2

Gaertner, W., & Schokkaert, E. (2012). *Empirical social choice: Questionnaire-experimental studies on distributive justice*. doi:https://doi.org/10.1017/CBO9781139012867

Gaertner, W., & Schwettmann, L. (2007). Equity, responsibility and the cultural dimension. *Economica*, *74*(296), 627–649. doi:https://doi.org/10.1111/j.1468-0335.2006.00563.x

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.

Genz, A. C., & Malik, A. A. (1980). Remarks on algorithm 006: An adaptive algorithm for numerical integration over an N-dimensional rectangular region. *Journal of Computational and Applied Mathematics*, *6*(4), 295–302. doi:10.1016/0771-050X(80)90039-X

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.

Gill, D., & Prowse, V. L. (2017). Using response times to measure strategic complexity and the value of thinking in games. *Available at SSRN 2902411*. doi:10.1007/s10683-015-9463-y

Gollwitzer, M., & van Prooijen, J.-W. (2016). Psychology of justice. In *Handbook of Social Justice Theory and Research* (pp. 61–82). doi:https://doi.org/10.1007/978-1-4939-3216-0_4

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.

Guo, M., Conitzer, V., & Reeves, D. M. (2009). Competitive repeated allocation without payments. In *International Workshop on Internet and Network Economics* (pp. 244–255). Springer. doi:https://doi.org/10.1007/978-3-642-10841-9_23

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.

Harris, M. R., & McGonigle, B. O. (1994). A model of transitive choice. *The Quarterly Journal of Experimental Psychology*, *47*(3), 319–348. doi:10.1080/14640749408401362

Harsanyi, J. C. (1975). Can the maximin principle serve as a basis for morality? A critique of John Rawls's theory. *American Political Science Review*, *69*(2), 594–606. doi:https://doi.org/10.2307/1959090

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

Herrero, C., Moreno-Ternero, J. D., & Ponti, G. (2010). On the adjudication of conflicting claims: An experimental study. *Social Choice and Welfare*, *34*(1), 145–179. doi:https://doi.org/10.1007/s00355-009-0398-0

Hirst, W., & Echterhoff, G. (2012). Remembering in conversations: The social sharing and reshaping of memories. *Psychology*, *63*(1), 55.

Hoffman, E., & Spitzer, M. L. (1985). Entitlements, rights, and fairness: An experimental examination of subjects' concepts of distributive justice. *Journal of Legal Studies*, *14*(2), 259–297. doi:https://doi.org/10.1086/467773

Hsu, M., Anen, C., & Quartz, S. (2008). The right and the good: Distributive justice and neural encoding of equity and efficiency. *Science*, *320*(5879), 1092–1095. doi:https://doi.org/10.1126/science.1153651

Hu, J., Lucas, C. G., Griffiths, T. L., & Xu, F. (2015). Preschoolers' understanding of graded preferences. *Cognitive Development*, *36*, 93–102. doi:10.1016/j.cogdev.2015.09.012

Huck, S., & Oechssler, J. (1999). The indirect evolutionary approach to explaining fair allocations. *Games and Economic Behavior*, *28*(1), 13–24. doi:https://doi.org/10.1006/game.1998.0691

Hurley, J., Buckley, N. J., Cuff, K., Giacomini, M., & Cameron, D. (2011). Judgments regarding the fair division of goods: The impact of verbal versus quantitative descriptions of alternative divisions. *Social Choice and Welfare*, *37*(2), 341–372. doi:https://doi.org/10.1007/s00355-010-0487-0

Jara-Ettinger, J. (2019). Theory of mind as inverse reinforcement learning. *Current Opinion in Behavioral Sciences*, *29*, 105–110. doi: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. doi: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. doi:10.1016/j.cognition.2017.06.017

Johansson-Stenman, O., Carlsson, F., & Daruvala, D. (2002). Measuring future grandparents' preferences for equality and relative standing. *Economic Journal*, *112*(479), 362–383. doi:https://doi.org/10.1111/1468-0297.00040

Jungeilges, J. A., & Theisen, T. (2008). A comparative study of equity judgements in Lithuania and Norway. *Journal of Socio-Economics*, *37*(3), 1090–1118. doi:https://doi.org/10.1016/j.socec.2007.04.002

Kalinowski, T., Narodytska, N., & Walsh, T. (2013). A social welfare optimal sequential allocation procedure. In *Twenty-Third International Joint Conference on Artificial Intelligence*.

Kash, I., Procaccia, A. D., & Shah, N. (2014). No agent left behind: Dynamic fair division of multiple resources. *Journal of Artificial Intelligence Research*, *51*, 579–603. doi:https://doi.org/10.1613/jair.4405

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. (2019). Revealed strength of preference: Inference from response times. *Judgment & Decision Making*, *14*(4).

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

Konow, J. (2000). Fair shares: Accountability and cognitive dissonance in allocation decisions. *American Economic Review*, *90*(4), 1072–1091. doi:https://doi.org/10.1257/aer.90.4.1072

Konow, J. (2001). Fair and square: The four sides of distributive justice. *Journal of Economic Behavior & Organization*, *46*(2), 137–164. doi:https://doi.org/10.1016/S0167-2681(01)00194-9

Konow, J. (2003). Which is the fairest one of all? A positive analysis of justice theories. *Journal of Economic Literature*, *41*(4), 1188–1239. doi:https://doi.org/10.1257/002205103771800013

Konow, J. (2009). Is fairness in the eye of the beholder? An impartial spectator analysis of justice. *Social Choice and Welfare*, *33*(1), 101–127. doi:https://doi.org/10.1007/s00355-008-0348-2

Konow, J., & Schwettmann, L. (2016). The Economics of justice. In *Handbook of Social Justice Theory and Research* (pp. 83–106). doi:https://doi.org/10.1007/978-1-4939-3216-0_5

Krajbich, I. (2019). Accounting for attention in sequential sampling models of decision making. *Current Opinion in Psychology*, *29*, 6–11. doi:10.1016/j.copsyc.2018.10.008

Krajbich, I., Armel, C., & Rangel, A. (2010). Visual fixations and the computation and comparison of value in simple choice. *Nature Neuroscience*, *13*(10), 1292. doi:10.1038/nn.2635

Krajbich, I., Bartling, B., Hare, T., & Fehr, E. (2015). Rethinking fast and slow based on a critique of reaction-time reverse inference. *Nature Communications*, *6*(1), 1–9. doi:10.1038/ncomms8455

Krajbich, I., Hare, T., Bartling, B., Morishima, Y., & Fehr, E. (2015). A common mechanism underlying food choice and social decisions. *PLoS Computational Biology*, *11*(10), e1004371. doi:10.1371/journal.pcbi.1004371

Krajbich, I., Oud, B., & Fehr, E. (2014). Benefits of neuroeconomic modeling: New policy interventions and predictors of preference. *American Economic Review*, *104*(5), 501–06. doi:10.1257/aer.104.5.501

Krajbich, I., & Rangel, A. (2011). Multialternative drift-diffusion model predicts the relationship between visual fixations and choice in value-based decisions. *Proceedings of the National Academy of Sciences*, *108*(33), 13852–13857. doi:10.1073/pnas.1101328108

Kuderer, M., Gulati, S., & Burgard, W. (2015). Learning driving styles for autonomous vehicles from demonstration. In *Robotics and Automation (ICRA), 2015 IEEE International Conference on* (pp. 2641–2646). IEEE. doi:https://doi.org/10.1109/ICRA.2015.7139555

Kurokawa, D., Procaccia, A. D., & Wang, J. (2016). When can the maximin share guarantee be guaranteed? In *AAAI* (Vol. 16, pp. 523–529).

Kuzmics, C., Palfrey, T., & Rogers, B. W. (2014). Symmetric play in repeated allocation games. *Journal of Economic Theory*, *154*, 25–67. doi:https://doi.org/10.1016/j.jet.2014.08.002

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, *4*, 863. doi:10.3389/fpsyg.2013.00863

Lerner, M. J. (1975). The justice motive in social behavior: Introduction. *Journal of Social Issues*, *31*(3), 1–19. doi:https://doi.org/10.1111/j.1540-4560.1975.tb00995.x

Leventhal, G. S. (1976). *Fairness in social relationships*. General Learning Press Morristown, NJ.

Levine, D. K. (1998). Modeling altruism and spitefulness in experiments. *Review of Economic Dynamics*, *1*(3), 593–622. doi:https://doi.org/10.1006/redy.1998.0023

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

Marion, S. B., & Thorley, C. (2016). A meta-analytic review of collaborative inhibition and post-collaborative memory: Testing the predictions of the retrieval strategy disruption hypothesis. *Psychological Bulletin*, *142*(11), 1141.

Marwell, G., & Ames, R. E. (1981). Economists free ride, does anyone else? Experiments on the provision of public goods, IV. *Journal of Public Economics*, *15*(3), 295–310. doi:https://doi.org/10.1016/0047-2727(81)90013-X

Maybery, M. T., Bain, J. D., & Halford, G. S. (1986). Information-processing demands of transitive inference. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *12*(4), 600. doi:10.1037/0278-7393.12.4.600

Meade, M., Nokes, T., & Morrow, D. (2009). Expertise promotes facilitation on a collaborative memory task. *Memory*, *17*(1), 39–48.

Milosavljevic, M., Malmaud, J., Huth, A., Koch, C., & Rangel, A. (2010). The drift diffusion model can account for the accuracy and reaction time of value-based choices under high and low time pressure. *Judgment and Decision Making*, *5*(6), 437. doi:10.2139/ssrn.1901533

Mitchell, G., Tetlock, P. E., Mellers, B. A., & Ordonez, L. D. (1993). Judgments of social justice: Compromises between equality and efficiency. *Journal of Personality and Social Psychology*, *65*(4), 629. doi:https://doi.org/10.1037/0022-3514.65.4.629

Moffatt, P. G. (2005). Stochastic choice and the allocation of cognitive effort. *Experimental Economics*, *8*(4), 369–388. doi:10.1007/s10683-005-5375-6

Moulin, H. [H.], Brandt, F., Conitzer, V., Endriss, U., Procaccia, A. D., & Lang, J. (2016). *Handbook of Computational Social Choice*. doi:https://doi.org/10.1017/CBO9781107446984

Moulin, H. [Hervé], & Stong, R. (2002). Fair queuing and other probabilistic allocation methods. *Mathematics of Operations Research*, *27*(1), 1–30. doi:https://doi.org/10.1287/moor.27.1.1.336

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). San Francisco, CA: Morgan Kaufmann.

Nord, E., Richardson, J., Street, A., Kuhse, H., & Singer, P. (1995). Who cares about cost? Does economic analysis impose or reflect social values? *Health Policy*, *34*(2), 79–94. doi:https://doi.org/10.1016/0168-8510(95)00751-D

Nowak, M. A., Page, K. M., & Sigmund, K. (2000). Fairness versus reason in the ultimatum game. *Science*, *289*(5485), 1773–1775. doi:https://doi.org/10.1126/science.289.5485.1773

Oleson, P. E. (2001). An experimental examination of alternative theories of distributive justice and economic fairness.

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.

Ordóñez, L. D., & Mellers, B. A. (1993). Trade-offs in fairness and preference judgments. doi:https://doi.org/10.1017/CBO9780511552069.008

Oud, B., Krajbich, I., Miller, K., Cheong, J. H., Botvinick, M., & Fehr, E. (2016). Irrational time allocation in decision-making. *Proceedings of the Royal Society B: Biological Sciences*, *283*(1822), 20151439. doi:10.1098/rspb.2015.1439

Overschelde, J. V., Rawson, K., & Dunlosky, J. (2004). Category norms: An updated and expanded version of the norms. *Journal of Memory and Language*, *50*(3), 289–335.

Pálvölgyi, D., Peters, H. J. M., & Vermeulen, A. J. (2010). A strategic approach to estate division problems with non-homogenous preferences. *METEOR Research Memorandum 10/036*. Maastricht.

Pelligra, V., & Stanca, L. (2013). To give or not to give? Equity, efficiency and altruistic behavior in an artefactual field experiment. *Journal of Socio-Economics*, *46*, 1–9. doi:https://doi.org/10.1016/j.socec.2013.05.015

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.

Polanìa, R., Krajbich, I., Grueschow, M., & Ruff, C. C. (2014). Neural oscillations and synchronization differentially support evidence accumulation in perceptual and value-based decision making. *Neuron*, *82*(3), 709–720. doi:10.1016/j.neuron.2014.03.014

Procaccia, A. D., & Wang, J. (2014). Fair enough: Guaranteeing approximate maximin shares. In *Proceedings of the Fifteenth ACM Conference on Economics and Computation* (pp. 675–692). ACM. doi:https://doi.org/10.1145/2600057.2602835

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

Rand, D. G., Greene, J. D., & Nowak, M. A. (2012). Spontaneous giving and calculated greed. *Nature*, *489*(7416), 427–430. doi:10.1038/nature11467

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*(2), 59. doi:10.1037/h0035486

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation*, *20*(4), 873–922. doi:10.1162/neco.2008.12-06-420

Ratcliff, R., Smith, P. L., Brown, S. D., & McKoon, G. (2016). Diffusion decision model: Current issues and history. *Trends in Cognitive Sciences*, *20*(4), 260–281. doi:10.1016/j.tics.2016.01.007

Rawls, J. (1971). A theory of justice. Belknap Press, Cambridge, Massachusetts.

Rawls, J. (1974). Some reasons for the maximin criterion. *American Economic Review*, *64*(2), 141–146.

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.

Rubinstein, A. (2007). Instinctive and cognitive reasoning: A study of response times. *The Economic Journal*, *117*(523), 1243–1259. doi:10.1111/j.1468-0297.2007.02081.x

Salles, R. M., & Barria, J. A. (2008). Lexicographic maximin optimisation for fair bandwidth allocation in computer networks. *European Journal of Operational Research*, *185*(2), 778–794. doi:https://doi.org/10.1016/j.ejor.2006.12.047

Schäfer, M., Haun, D., & Tomasello, M. (2015). Fair is not fair everywhere. *Psychological Science*, 1252–1260. doi:https://doi.org/10.1177/0956797615586188

Schokkaert, E., & Devooght, K. (2003). Responsibility-sensitive fair compensation in different cultures. *Social Choice and Welfare*, *21*(2), 207–242. doi:https://doi.org/10.1007/s00355-003-0257-3

Schokkaert, E., & Lagrou, L. (1983). An empirical approach to distributive justice. *Journal of Public Economics*, *21*(1), 33–52. doi:https://doi.org/10.1016/0047-2727(83)90072-5

Schokkaert, E., & Overlaet, B. (1989). Moral intuitions and economic models of distributive justice. *Social Choice and Welfare*, *6*(1), 19–31. doi:https://doi.org/10.1007/BF00433360

Schwettmann, L. (2009). *Trading off competing allocation principles: Theoretical approaches and empirical investigations*. Peter Lang.

Schwettmann, L. (2012). Competing allocation principles: Time for compromise? *Theory and Decision*, *73*(3), 357–380. doi:https://doi.org/10.1007/s11238-011-9289-9

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.

Shadlen, M. N., & Shohamy, D. (2016). Decision making and sequential sampling from memory. *Neuron*, *90*(5), 927–939. doi:10.1016/j.neuron.2016.04.036

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.

Shevlin, B., & Krajbich, I. (2020). Attention as a source of variability in decision-making: Accounting for overall-value effects with diffusion models. *PsyArXiv*. doi:10.31234/osf.io/rewtq

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

Skitka, L. J., & Tetlock, P. E. (1992). Allocating scarce resources: A contingency model of distributive justice. *Journal of Experimental Social Psychology*, *28*(6), 491–522. doi:https://doi.org/10.1016/0022-1031(92)90043-J

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.

Spiliopoulos, L., & Ortmann, A. (2018). The BCD of response time analysis in experimental economics. *Experimental Economics*, *21*(2), 383–433. doi:10.1007/s10683-017-9528-1

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

Thayer, E. S., & Collyer, C. E. (1978). The development of transitive inference: A review of recent approaches. *Psychological Bulletin*, *85*(6), 1327. doi:10.1037/0033-2909.85

Thomson, W. (2003). Axiomatic and game-theoretic analysis of bankruptcy and taxation problems: A survey. *Mathematical Social Sciences*, *45*(3), 249–297. doi:https://doi.org/10.1016/S0165-4896(02)00070-7

Thorley, C., & Dewhurst, S. (2007). Collaborative false recall in the DRM procedure: Effects of group size and group pressure. *European Journal of Cognitive Psychology*, *19*(6), 867–881.

Towal, R. B., Mormann, M., & Koch, C. (2013). Simultaneous modeling of visual saliency and value computation improves predictions of economic choice. *Proceedings of the National Academy of Sciences*, *110*(40), E3858–E3867. doi:10.1073/pnas.1304429110

Traub, S., Seidl, C., Schmidt, U., & Levati, M. V. (2005). Friedman, Harsanyi, Rawls, Boulding–or somebody else? An experimental investigation of distributive justice. *Social Choice and Welfare*, *24*(2), 283–309. doi:https://doi.org/10.1007/s00355-003-0303-1

Walsh, T. (2011). Online cake cutting. In *International Conference on Algorithmic Decision Theory* (pp. 292–305). Springer. doi:https://doi.org/10.1007/978-3-642-24873-3_22

Walsh, T. (2015). Challenges in resource and cost allocation. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

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. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1–6).

Wegner, D., Erber, R., & Raymond, P. (1991). Transactive memory in close relationships. *Journal of personality and social psychology*, *61*(6), 923.

Wilcox, N. T. (1993). Lottery choice: Incentives, complexity and decision time. *The Economic Journal*, *103*(421), 1397–1417. doi:10.2307/2234473

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

Wittig, M., Jensen, K., & Tomasello, M. (2013). Five-year-olds understand fair as equal in a mini-ultimatum game. *Journal of Experimental Child Psychology*, *116*(2), 324–337. doi:https://doi.org/10.1016/j.jecp.2013.06.004

Yaari, M. E., & Bar-Hillel, M. (1984). On dividing justly. *Social choice and welfare*, *1*(1), 1–24. doi: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*. arXiv: 1911.03020. Retrieved from http://arxiv.org/abs/1911.03020

Zhao, W. J., Diederich, A., Trueblood, J. S., & Bhatia, S. (2019). Automatic biases in intertemporal choice. *Psychonomic Bulletin & Review*, *26*(2), 661–668. doi:doi.org/10.3758/s13423-019-01579-9

Ziebart, B. D., Maas, A. L., Bagnell, J. A., & Dey, A. K. (2008). Maximum entropy inverse reinforcement learning. In *AAAI* (Vol. 8, pp. 1433–1438). Chicago, IL, USA.

Appendix A

Supplemental Material for Chapter 1

These Supplemental Materials contain additional details describing the model from Luhmann and Rajaram (2015).

A.1 Agent-Based Model: Methods and Results

In the model described by Luhmann and Rajaram (2015), agents encode N items (words), where N=40. Agents have two representations. The first is an activation vector $\mathbf{A}$ of length N. Each entry $\mathbf{A}_j$ gives the probability that the jth item will be retrieved. The second representation is an inter-item association matrix $\mathbf{S}$ of size $N \times N$. Each entry $\mathbf{S}_{ij}$ gives the jth item's (possibly asymmetric) association with the ith item. This matrix would normally contain agents' prior knowledge about word associations; however, Luhmann and Rajaram (2015) assigned values of $\mathbf{S}$ randomly between -2 and 2 to reflect agnosticism about the semantic relationships between words. (Neither the activation vector $\mathbf{A}$ nor association matrix $\mathbf{S}$ were used in our empirical studies because they were not applicable: we did not have access to participants' internal memory representations, nor did we test the theory that participants' updated these representations based on (randomly-selected or otherwise) semantic associations between words.)

The agents are exposed to items one by one, encoding them. The first step in encoding an item is

to reduce the activation of the maximally active item in vector $\mathbf{A}$, where β is the learning rate:

$$\Delta \mathbf{A}_{\text{max}} = -\beta \mathbf{A}_{\text{max}} \,. \tag{A.1}$$

Next, the agent reduces the activations of items semantically associated with the maximally active item:

$$\Delta \mathbf{A}_j = -\beta \mathbf{S}_{j,\text{max}} \mathbf{A}_j \,. \tag{A.2}$$

Finally, the agent increases the activation of the to-be-encoded item, with α acting as the learning rate:

$$\Delta \mathbf{A}_i = \alpha \big[1 - \mathbf{A}_i\big] \,. \tag{A.3}$$

The activation vector is then normalized to ensure that its entries can be interpreted as probabilities: $\sum_i \mathbf{A}_i = 1$.

During recall, an agent retrieves (and "orally states") an item. Agents take turns retrieving items, and on each turn, an agent retrieves an item with probability γ. Items are retrieved according to $\mathbf{A}$, such that items with higher activations are more likely to be retrieved.

The act of retrieving an item modifies the activation vector for both the retriever and the other members of the group. First, the agent decreases activation of items semantically associated with the retrieved item, in line with the theory of retrieval disruption:

$$\Delta \mathbf{A}_j = \beta \mathbf{S}_{ji} \mathbf{A}_j \,. \tag{A.4}$$

Next, if the ith item is not the maximally activated item, the agent reduces the activation of the maximally active item according to Equation 1, and the activations of items semantically associated with the maximally active item according to Equation 2. Item i is then encoded according to Equation 3. $\mathbf{A}$ is then normalized such that $\sum_i \mathbf{A}_i = 1$. Just as the retrieving agent encodes the item after retrieving it, "listening" agents also then encode the item according to the encoding process described previously. Luhmann and Rajaram (2015) used the following parameter settings: α=0.2, β=0.05, and γ=0.75.

Model predictions were generated by presenting agents with wordlists and then having agents recall words via the described procedure. When an agent generated a word, it was shared with every other agent in the network. Agents participated in 20 rounds of retrieval within each simulation. A total of 1000 simulations (comparing 1000 collaborative and 1000 nominal results) were run for each group size.

This model makes implicit assumptions through its paradigm choices, for example by using a

turn-based paradigm with words "read aloud" to agents. In our empirical experiments, we tried to match these model choices as closely as possible, to best compare our behavioral results to model predictions. However, an exception is that in the modeling work, agents were allowed to submit any word that they had not previously retrieved, whereas in the behavioral work, participants were not allowed to recall words that they or any other group members had previously recalled.

A.1.1 Model Comparison: 40 words vs. 60 words

In Luhmann and Rajaram (2015), the authors use the following parameter settings to generate their model predictions: 40 words, α=0.2, β=0.05, γ=0.75, number of rounds=20, number of simulations=1000. Using these settings, they predict the following set of results:

> "As groups grew from 2 to 7 agents, collaborative inhibition increased (a finding that replicates and extends those of Basden et al., 2000, and Thorley & Dewhurst, 2007). In this range, the performance of both collaborative and nominal groups increased steadily. However, each additional group member conferred a much larger benefit to nominal groups than to collaborative groups. This effect of group size was probably driven by the relative balance between the facilitative effects offered by collaboration (i.e., more agents increased the probability that the group would retrieve a given item) and the detrimental effects of retrieval disruption (i.e., more collaborators meant more opportunities to be disrupted). Collaborative inhibition decreased as group size increased beyond 7. This effect was driven by the fact that nominal groups reached ceiling far earlier than the collaborative groups. Nonetheless, the continued deficiencies exhibited by even large collaborative groups suggest that the cognitive factors that hurt retrieval diversity in small groups could not easily be overcome by the addition of group members."

In our work, we use the code from Luhmann and Rajaram (2015) to generate model predictions with the number of presented words set at 60 rather than 40. As such, we do not expect to see collaborative inhibition peaking at a group size of 7, since this is an arbitrary number responsive to the parameter settings. However, we do expect to see the general trends Luhmann and Rajaram (2015) describe: collaborative recall increasing with group size, until nominal recall reaches ceiling performance as the disruption of idiosyncratic recall strategies is compensated for by sheer group size. The model will predict that from this point collaborative inhibition begins to decrease, since collaborative inhibition is the difference between nominal and collaborative recall. With our parameter settings, nominal ceiling performance occurs around group size 16. Note that we use identical parameter settings to Luhmann and Rajaram (2015) except for using 60 words, chosen because pilot

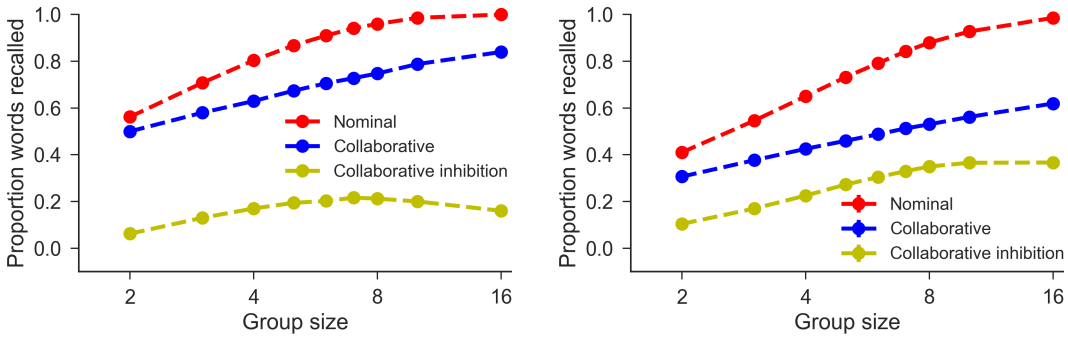


Figure A.1: Comparison of model predictions for 40 vs. 60 recalled words. (**Left**) Model predictions reproduced from Luhmann and Rajaram (2015), using parameter settings: 40 words, α=.2, β=.05, γ=.75, number of rounds=20, number of simulations=1000. Nominal recall (red/grey), collaborative recall (blue/black), and the subtraction of the two, collaborative inhibition (yellow/light-grey) are shown. (**Right**) Model predictions produced using code from Luhmann and Rajaram (2015), using 60 words but otherwise identical parameters settings.

studies indicated participants were nearing ceiling performance at small group sizes. Here we provide a comparison of the figure originally used in Luhmann and Rajaram (2015), using 40 words, compared to the model predictions generated using 60 recalled words. (Figure A.1).

Appendix B

Supplemental Material for Chapter 3

B.1 Experiment 2, Additional Analysis

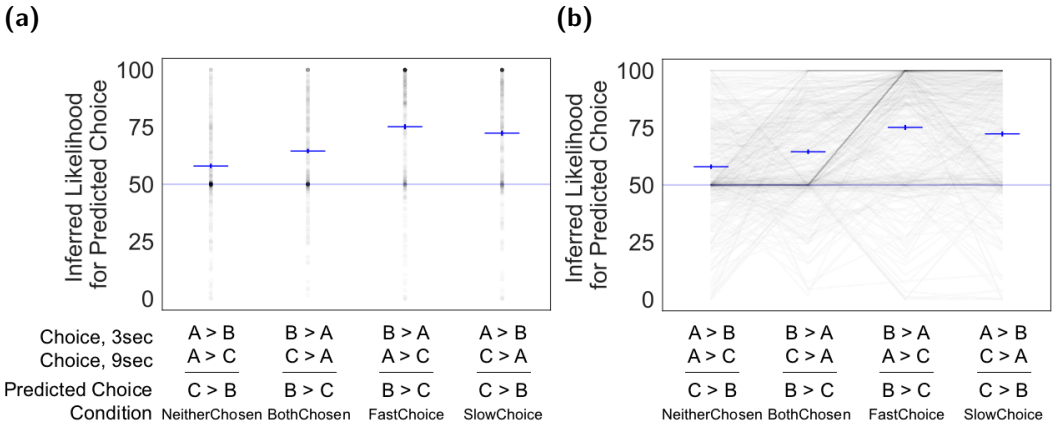


Figure B.1: Experiment 2 Results, Additional Analysis. (a) Main Figure (see Figure 3.6). (b) Individual (Connected) Participant Responses. Note the prevalence of responses following a "50 - 50 - 100 - 100" pattern for the conditions "NeitherChosen - BothChosen - FastChoice - SlowChoice."

Future work may also explore some quirks of the empirical results. In Figure 3.6, the results indicate that while participants did not often predict the non-predicted choice, very many participants chose 50 for the timing-inference based conditions, and 100 for the control conditions. (These num-

bers indicate participants' inferred likelihood for the predicted choice; the critical question participants were asked was: "This person is now offered a choice between items B and C. What choice do you think they'd make, and how likely do you think it is that they'd make that choice?") This pattern indicates that participants were not using response times when making their choices, and that choices alone could make them estimate 100% likelihoods for the decision-maker's choice on an unseen choice. Figure B.1 suggests that a number of participants answered in this "50-50-100-100" sequence (for conditions "NeitherChosen-BothChosen-FastChoice-SlowChoice"), implying that if such responses had been discouraged the effect observed here may have been even stronger.

B.2 Experiment 3, Excluded Participants

The analysis was identical to Experiment 3 except that participants were excluded if they answered more than two out of eight manipulation check questions incorrectly, as was defined in the preregistration. Of 481 recruited participants, 212 met this exclusion criteria, so 269 participants were included. These results were similar to the Experiment 3 results with all participants included, and are described below (Figure B.2).

From our linear mixed model, the estimate (beta parameter) for the main effect of response time was 6.58 (95%CI=[4.53,8.62], t(1880.0)=6.31, p=4e-10, t-tests using Satterthwaite's method, η_p^2=0.02). The estimate for the main effect of threshold was 19.59 (95%CI=[17.54,21.63], t(1880.0)=18.70, p<2e-16, t-tests using Satterthwaite's method, η_p^2=0.16). The estimate for the interaction effect of response time and threshold was 3.65 (95%CI=[-0.43,7.74], t(1880.0)=1.75, p=0.08, t-tests using Satterthwaite's method, η_p^2=0.002): not a significant effect. The estimated power of the interaction effect for α=0.01 was 20.00% (95%CI=[14.69,26.22]) using a Type-III F-test from the R package "car," generated via 200 simulations with the R package "simr." (The estimate for the fixed effect intercept was 57.37, 95%CI=[55.67,59.08], t(268.0)=66.15, p<2e-16, t-tests using Satterthwaite's method, η_p^2=0.94.)

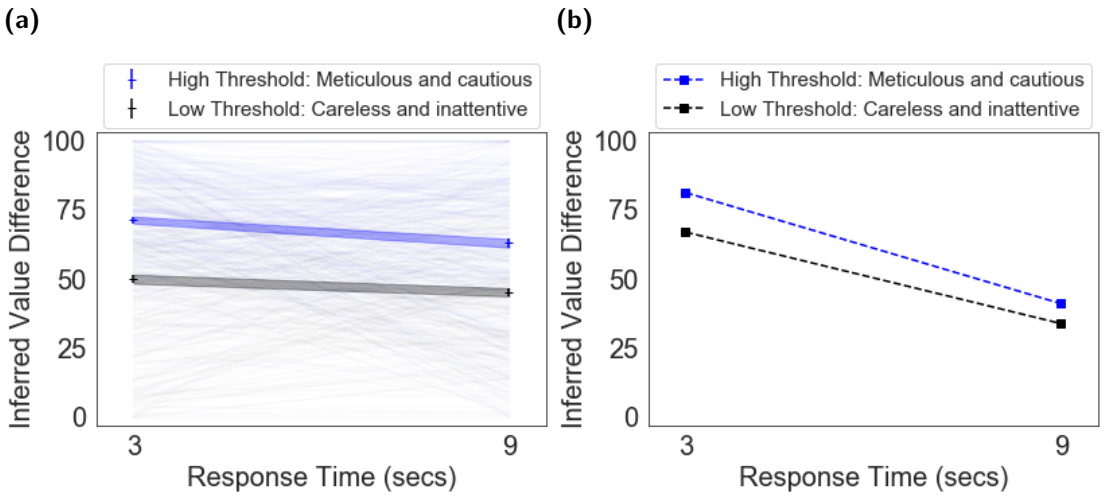
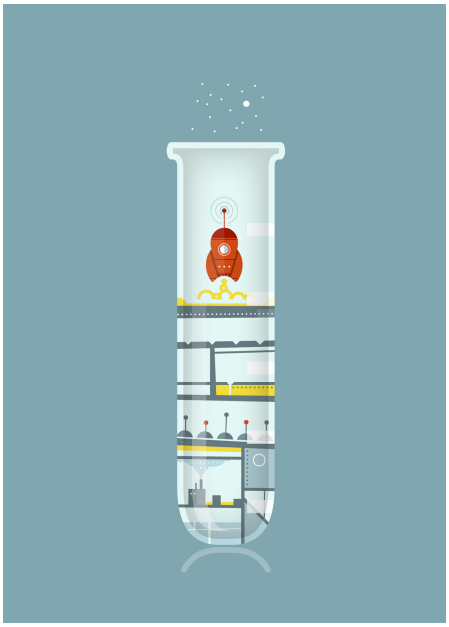


Figure B.2: Experiment 3 Results, with Excluded Participants. (a) Experimental Results. The mean $\pm$ SE of the inferred value difference, n=269, is shown for the high threshold ("cautious," blue) and the low threshold ("careless," black) conditions, for each of the 3 and 9 second response time conditions. Responses are connected by threshold condition for emphasis. Individual participants' averaged values for each threshold/time pair are also shown and connected. 0 represents the participant feeling that the decision-maker was neutral between items, 50 that the decision-maker moderately preferred their chosen item, and 100 that the decision-maker strongly preferred their chosen item. (b) Model Predictions.



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.