UNIVERSITY OF CALIFORNIA SAN DIEGO

Building Mental Models of Others Over Repeated Interactions

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy

in

Experimental Psychology

by

Erik Brockbank

Committee in charge:

Professor Judith Fan, Co-Chair
Professor Edward Vul, Co-Chair
Professor Marcelo Mattar
Professor Isabel Trevino
Professor Caren Walker

2023

The Dissertation of Erik Brockbank is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2023

DEDICATION

To my parents, for letting me stay home from school whenever I wanted to play with Legos or read Calvin and Hobbes in the backyard instead.

I think I've made up for all the absences by now...

EPIGRAPH

"The scientific enterprise as a whole does from time to time prove useful, open up new territory, display order, and test long-accepted belief. Nevertheless, *the individual* engaged on a normal research problem *is almost never doing any one of these things.* Once engaged, his motivation is of a rather different sort. What then challenges him is the conviction that, if only he is skilful enough, he will succeed in solving a puzzle that no one before has solved or solved so well. Many of the greatest scientific minds have devoted all of their professional attention to demanding puzzles of this sort. On most occasions any particular field of specialization offers nothing else to do, a fact that makes it no less fascinating to the proper sort of addict."

Kuhn (1962)

"[T]he key is not the *stuff* out of which brains are made, but the *patterns* that can come to exist inside the stuff of a brain."

Hofstadter (1979)

"Tell me how all this, and love too, will ruin us.
These, our bodies, possessed by light.
Tell me we'll never get used to it."

Siken (2005)

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGEMENTS

I wish to extend a deep and heartfelt gratitude to the members of my committee, in particular to **Judy Fan**, **Ed Vul**, and **Caren Walker** for their support and energy throughout my time in graduate school. Caren welcomed me into her lab from day one, administered the red pen treatment with clinical precision to countless paper drafts in the years that followed, and has advocated for me every step of the way. I learned so much from her tireless insistence on *getting it right* in literature reviews and discussion sections in particular; these were lessons not just about how to write well, but how to do justice to the ideas and effort of the broader community. Beyond this, she was a caring and empathetic listener throughout the many bumps along the way, for which I am ever appreciative. To Ed I owe so much of my intellectual development over the past five years. In particular, his emphasis on simplicity and his ability to cut to the bottom of theoretical arguments, computational methods, and behavioral experiments equally is something I hope to take with me everywhere. On top of this, the joy, curiosity and humor that he brings to these tasks kept the whole process enjoyable even in the most challenging moments. I am so grateful that he was willing to take a flyer on a graduate school applicant five years out of school, with little relevant research experience, who wrote in their cover letter and insisted in interviews that they wanted to study "learning." It was no doubt a daunting task to shepherd that person through the first several years of graduate school, but somehow Ed did it with the utmost grace. Finally, to Judy, it feels like a very long time ago that she sat next to me at a diner on the last day of CogSci and asked what I was going to work on in grad school, then listened patiently as I spat out half-baked nonsense about unlocking the secrets of cognition using entropy in repeated rock, paper, scissors games. Since then, the outpouring of investment in my work, career, and interests has felt often undeserved but always deeply appreciated. The way she approaches this job is inspiring and I'm grateful to be along for the ride; I still have so much to learn about how to tell the story right, but the progress I've made so far has been largely on Judy's shoulders.

Chapter 1, in full, is a reprint of material as it appears in Brockbank, E., & Vul, E. (2021). Formalizing opponent modeling with the rock, paper, scissors game. *Games*, *12*(3), 70. An earlier version of the project was published as Brockbank, E., & Vul, E. (2020). Recursive adversarial reasoning in the rock, paper, scissors game. In S. Denison., M. Mack, Y. Xu, & B.C. Armstrong (Eds.), *Proceedings of the 42nd Annual Conference of the Cognitive Science Society* (pp. 1015-1021). Cognitive Science Society. The dissertation author was the primary investigator and author of this material.

Chapter 2, in full, has been submitted for publication of the material and is currently in revision as it may appear in *Cognitive Psychology*. An earlier version of the project was published as Brockbank, E., & Vul, E. (2021). Humans fail to outwit adaptive rock, paper, scissors opponents. In T. Fitch, C. Lamm, H. Leder, & K. Teßmar-Raible (Eds.), *Proceedings of the 43rd Annual Meeting of the Cognitive Science Society* (pp. 1740-1746). Cognitive Science Society. The dissertation author was the primary investigator and author of this material.

Chapter 3, in full, is currently being prepared for submission for publication of the material. An earlier version of the project was published as Brockbank, E., Wang, H., Yang, J., Mirchandani, S., Bıyık, E., Sadigh, D., & Fan, J. E.. (2022). How do people incorporate advice from artificial agents when making physical judgments? In J. Culbertson, A. Perfors, H. Rabagliati, & V. Ramenzoni (Eds.), *Proceedings of the 44th Annual Meeting of the Cognitive Science Society* (pp. 1469-1475). Cognitive Science Society. The dissertation author was the primary investigator and author of this material.

VITA

2013	Bachelor of Science, Symbolic Systems, Stanford University
2014	Master of Science, Symbolic Systems, Stanford University
2023	Doctor of Philosophy, Experimental Psychology, University of California San Diego

PUBLICATIONS

Brockbank, E., Holdaway, C., Acosta-Kane, D., & Vul, E. (2023). Sampling data, beliefs, and actions. In K. Fiedler, P. Juslin, & J. Denrell (Eds.), *Sampling in judgment and decision making* (pp. 513–548). Cambridge University Press.

Brockbank, E., Lombrozo, T., Gopnik, A., & Walker, C. M. (2023). Ask me why, don't tell me why: Asking children for explanations facilitates relational thinking. *Developmental Science*, *26*(1), e13274.

Brockbank, E., Barner, D., & Vul, E. (2022). Ongoing dynamic calibration produces unstable number estimates. *Journal of Experimental Psychology: General*, *151*(9), 2092.

Brockbank, E., & Walker, C. M. (2022). Explanation impacts hypothesis generation, but not evaluation, during learning. *Cognition*, *225*, 105100.

Brockbank, E., *Wang, H., Yang, J., Mirchandani, S., Bıyık, E., Sadigh, D., & Fan, J. E. (2022). How do people incorporate advice from artificial agents when making physical judgments? In J. Culbertson, H. Rabagliati, V. Ramenzoni, & A. Perfors (Eds.), *Proceedings of the 44th annual conference of the cognitive science society* (pp. 1375–1381). Cognitive Science Society.

Schneider, R. M., Brockbank, E., Feiman, R., & Barner, D. (2022). Counting and the ontogenetic origins of exact equality. *Cognition*, *218*, 104952.

Brockbank, E., & Vul, E. (2021a). Formalizing opponent modeling with the rock, paper, scissors game. *Games*, *12*(3), 70.

Brockbank, E., & Vul, E. (2021b). Humans fail to outwit adaptive rock, paper, scissors opponents. In T. Fitch, C. Lamm, H. Leder, & K. Teßmar-Raible (Eds.), *Proceedings of the 43rd annual conference of the cognitive science society* (pp. 1740–1746). Cognitive Science Society.

DeStefano, I., *Oey, L. A., Brockbank, E., & Vul, E. (2021). Integration by parts: Collaboration and topic structure in the cogsci community. *Topics in Cognitive Science*, *13*(2), 399–413.

Brockbank, E., & Vul, E. (2020). Recursive adversarial reasoning in the rock, paper, scissors game. In S. Denison, M. L. Mack, Y. Xu, & B. C. Armstrong (Eds.), *Proceedings of the 42nd annual conference of the cognitive science society* (pp. 1015–1021). Cognitive Science Society.

Brockbank, E., & Walker, C. M. (2020). Explanation supports hypothesis generation in learning. In S. Denison, M. L. Mack, Y. Xu, & B. C. Armstrong (Eds.), *Proceedings of the 42nd annual conference of the cognitive science society* (pp. 74–80). Cognitive Science Society.

Oey, L., *DeStefano, I., Brockbank, E., & Vul, E. (2020). Formalizing interdisciplinary collaboration in the cogsci community. In S. Denison, M. L. Mack, Y. Xu, & B. C. Armstrong (Eds.), *Proceedings of the 42nd annual conference of the cognitive science society* (pp. 474–480). Cognitive Science Society.

Brockbank, E., & Vul, E. (2019). Mapping visual features onto numbers. In C. Freksa, A. Goel, & C. Seifert (Eds.), *Proceedings of the 41st annual conference of the cognitive science society* (pp. 1443–1449). Cognitive Science Society.

ABSTRACT OF THE DISSERTATION

Building Mental Models of Others Over Repeated Interactions

by

Erik Brockbank

Doctor of Philosophy in Experimental Psychology

University of California San Diego, 2023

Professor Judith Fan, Co-Chair
Professor Edward Vul, Co-Chair

Human interaction relies on the ability to form accurate internal models of other people. *What is the structure of our mental representations of others?* Existing theories in psychology broadly fall into two classes: those which view people as constructing rich generative models of those around us, and those which argue for more simplified predictive representations based on past behavior. In this dissertation, I explore the conditions under which people employ different representations of others and the constraints they face in each case. My work probes dyadic behavior across repeated interactions, thereby exposing the precise structure of the representations that people form in diverse settings.

In Chapter 1, I begin by investigating how people develop predictive models of others based purely on simple, sequential patterns in their previous actions. I present evidence that in *mixed strategy equilibrium* (MSE) games, people acquire an adaptive model of their opponent over many interactions and argue that behavior in such games offers a novel perspective on people's opponent modeling. In Chapter 2, I present two studies characterizing the basis of people's opponent modeling in MSE games and exploring the scope of this ability. Results suggest that people show substantial limitations in their capacity to develop predictive models of others using patterns in their behavior alone. In light of these findings, Chapter 3 explores the process by which people develop more abstract and sophisticated representations of others in domains where they have rich mental models of their own. Specifically, this work focuses on how people incorporate the *competence* of another agent into collaborative interactions in a physical task. I first show that people infer latent and dynamic properties of others' behavior in this setting; in a second study, I show that such inferences extend to features of their collaborator's internal model of the task. Broadly, this work suggests that our representations of others can take on surprisingly diverse forms but their complexity is heavily context-dependent. I conclude with a discussion of future directions aimed at understanding the structure of people's representations of others and how they calibrate these representations to the context at hand.

Introduction

0.1 Overview

What form do our mental representations of others take? The ability to leverage sophisticated yet flexible internal models of those around us to predict and explain their behavior is central to a range of uniquely human activities, from coordination and complex planning to communication and forming meaningful relationships. Researchers in psychology and artificial intelligence have long been interested in understanding human *intuitive psychology* (Heider, 1958), yet only recently has the scale of modern computation and developments in cognitive modeling allowed for robust and fine-grained tests of distinct theories about how we represent the actions and mental states of others (Alanqary et al., 2021; Baker et al., 2017). The work in this dissertation explores this exciting intersection of computational methods and social psychology to understand how we represent those around us in diverse contexts. I investigate dyadic behavior over repeated interactions in both adversarial and cooperative settings to better understand how we develop internal models of those around us and what these internal models are made of; this work exposes the precise structure of our internal models of others in these settings, as well as the constraints people face in forming such representations. The results point the way towards future work aimed at providing a broad account of the representations underlying our intuitive psychology.

0.2 "Behaviorist" and "cognitivist" intuitive psychology

Over the last 100 years, the field of experimental psychology underwent a dramatic shift in methodology and theoretical stance, often referred to as the "Cognitive Revolution" (Mandler, 2002; G. A. Miller, 2003). At stake in this transition to what we now consider modern psychology was what kind of *mental representations* could be attributed to people on the basis of their behavior. The predominant paradigm prior to the Cognitive Revolution, *behaviorism*, was primarily concerned with explanations for human and animal behavior based on learned associations between stimuli and rewards (Skinner, 1938; Thorndike, 1911; Watson, 1924)—this account reflects a notable sparsity of mental representation. In contrast, proponents of the new *cognitivist* view argued that predominant features of human psychology such as syntax (Chomsky, 1957), working memory limitations (G. A. Miller, 1956), and problem solving (Newell et al., 1958) could not be explained without recourse to cognitive operations or mental states that extended beyond the purview of learned associations. This emphasis on unobservable mental states and operations as causes of behavior was central to the subsequent rise of *cognitive science* and is now a broadly accepted view among psychologists (Gardner, 1987; Núñez et al., 2019). However, the role of reinforcement learning and other computational approaches which owe their origins to behaviorism in modern cognitive neuroscience (Dayan & Niv, 2008) and artificial intelligence (Sutton & Barto, 1998) underscores the fact that foundational contributions from behaviorism remain relevant even today.

The tension between behaviorism and cognitivism centered on the empirical psychology of humans (and other animals) interacting with their environment. However, it offers an instructive framework with which to view current research into our *intuitive psychology*. What kind of mental states and representations do we attribute to *others* as the cause of their behavior? At one extreme, we might imagine other people the way behaviorists imagined people; our internal model of those around us could make no use of mental representations at all, instead viewing others as merely exhibiting learned associations between stimulus and reward. For example, if we see a co-worker arrive at the office with a jacket, we might infer that they brought a jacket because it was raining at their house when they left, or simply because they always bring a jacket to work (see Figure 1). At the other end of the spectrum, our internal model of others might align with contemporary psychological accounts of individual planning and decision-making, where actions reflect the joint influence of a host of interconnected mental states: goals, beliefs, and subjective costs and rewards (Callaway et al., 2022; M. K. Ho et al., 2022; Zhi-Xuan et al., 2020). On this view, when a co-worker arrives at the office with a jacket, we might infer that they *believe* it will rain at some point today, regardless of what the weather was like when they left (Figure 1). Finally, people's internal model of others may fluctuate between these two, sometimes relying on more simplistic *behaviorist* representations of others which require minimal mental state reasoning, and other times leveraging a more *cognitivist* framework which permits rich inferences about the mental states that produce others' behavior.

In short, just as behaviorism and modern cognitive psychology offer credible but distinct accounts of how an agent might learn and choose actions in their environment (each with strong formal and theoretical underpinnings), they can just as equally serve to anchor theorizing about our *internal models of other decision-makers*. And just as individual behavior might at times reflect more of one model or the other (Kahneman, 2011), we can just as easily ask under what conditions people's representation of others reflects a more behaviorist or cognitivist intuitive psychology. In fact, as I discuss below, this contrast between behaviorism and cognitivism may be even *more* relevant in theorizing about how we represent *others*, since reasoning about other minds poses a costly and challenging computational task above and beyond one's own planning and decision-making (Jara-Ettinger, 2019), which could at times tip the scales more towards behaviorist representations of those around us.

The work in this dissertation has been strongly influenced by the perspective that our representations of others may be structured in ways similar to competing theoretical accounts of individual reasoning from the recent history of psychology. However, it's important to note that this is not the only way of distinguishing between distinct models of others' behavior which propose differing amounts of underlying mental representation or planning. For example, the behaviorist-cognitivist distinction closely mirrors that of *model-free* and *model-based* reinforcement learning (Sutton & Barto, 1998). Such alternatives can likely be employed interchangeably in what follows, though the model-free/model-based structure is traditionally used in the context of *Markov Decision Processes* (Bellman, 1957), whereas the current analysis is meant to extend to reasoning about others in a broader range of settings. In addition, this conceptualization has been used in existing work; Bigelow and Ullman (2022) for example describe people's evaluations of other agents as reflecting "intuitive behaviorism" and "intuitive cognitivism". My own thoughts about this were developed independently, though the underlying concepts are likely similar.

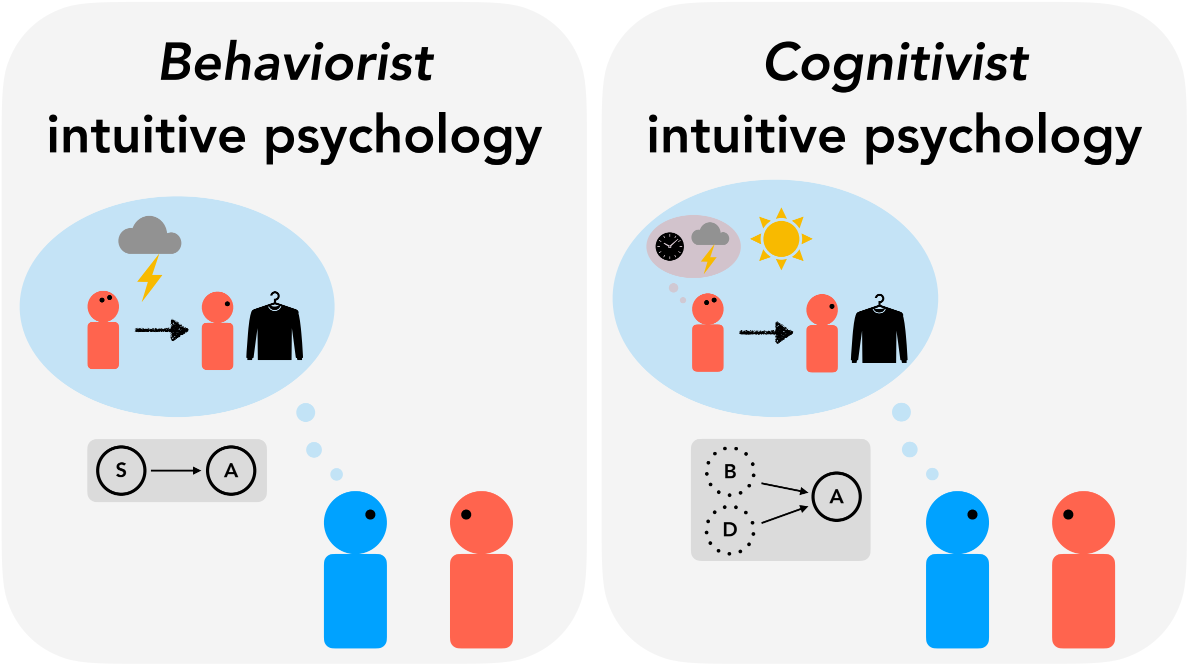


Figure 1. Theoretical framework for intuitive representations of others. At left, *behaviorist* intuitive psychology models represent others' actions as simple learned associations between stimulus (S) and action (A). At right, *cognitivist* models represent others' actions (A) as resulting from mental states like beliefs (B) and desires (D).

In what follows, I review prior findings examining *how our representations of others are structured*, with particular attention to what these results tell us about *behaviorist* and *cognitivist* mental models. I describe questions that remain unanswered in this literature and which the work in this dissertation attempts to address. Finally, I offer a brief outline of the results from my work and how this may support a broader and more comprehensive account of how we represent others across diverse settings and contexts.

0.3 Children are "born cognitivists"

Where does our intuitive psychology come from, and does its ontogeny show any sort of adherence to cognitivist or behaviorist mental models? An extensive body of work has sought to pin down the early structure and developmental origins of our *theory of mind* (ToM) (Premack & Woodruff, 1978); broadly, this is the idea that others have mental states like beliefs and desires which play a causal role in their subsequent actions. Though adults can and often do use this cognitivist model of others (see below), its development in children is typically slow and piecewise (Goodman et al., 2006; Gopnik & Wellman, 1992; Onishi & Baillargeon, 2005; Wellman et al., 2001).

Children's theory of mind is primarily studied with tasks that probe whether participants can recognize when another person has a *false belief* (Wimmer & Perner, 1983). The notion that another's beliefs can be subjective and inconsistent with our own and further that people's actions will reflect this false representation rather than our own knowledge or perceptual experience is a defining feature of beliefs as *mental states that uniquely guide behavior*. Critically, this understanding differentiates young children from adults; while adults have no trouble predicting how somebody with a false belief will behave in canonical false belief tasks, children undergo a dramatic developmental shift in this ability (Onishi & Baillargeon, 2005; Wellman et al., 2001). Gopnik and Wellman (1992) describe this developmental change as a gradual *theory revision* process between around two and five years of age, in which children's model of others' mental states proceeds from being largely restricted to desires and perceptions, to one in which others are understood to have beliefs but these are treated as veridical reflections of the world, to one in which beliefs are subjective and otherwise resemble the adult concept of a belief. This account, which describes an increasingly complex understanding of others' mental states from the get-go, leaves little room for the possibility of a *behaviorist theory of mind*, even in development; the authors address this question explicitly (emphasis added):

> The 2-year-old is clearly a mentalist and not a behaviorist. Indeed, **it seems unlikely to us that there is ever a time when normal children are behaviorists**... It seems plausible that mentalism is the starting state of psychological knowledge (p. 150).

Thus, while the historical progression from behaviorist to cognitivist theories offers an intuitive prediction about human development, it's not clear that children's increasingly sophisticated intuitive psychology follows this trajectory. Instead, it may be that the ability to conceptualize of others as behaving according to simple learned associations like *habits* or *rules* is only available later in development, or that the bias to attribute behavior to mental states must be overcome with experience.

0.4 Cognitivist intuitive psychology permits complex mental state inferences and behavior

Though decades of research has explored adults' tendency to infer goals and desires from others' behavior—even when such behavior provides very sparse visual cues to agency (Heider & Simmel, 1944)—it is only recently that inroads have been made on the computational underpinnings of this process (Baker et al., 2017). This work sheds light on the structure of adults' cognitivist representations of others, and has spurred corresponding investigations into these same representations in children. Recent computational accounts of how we reason about others rely on *inverse planning* as a model of inferring the mental states that caused another agent's actions (Baker et al., 2017; Baker et al., 2009; Jara-Ettinger et al., 2018; Ullman et al., 2009). This work proposes that we tend to view others as *rational planners*; this assumption allows reasoners to back out the kinds of goals and beliefs that best account for others' actions. Formally, these results rely on tasks in which other agents' behavior in simple grid worlds can be modeled using Markov Decision Processes (Bellman, 1957); reasoning about those agents' mental states is then achieved through Bayesian inference over parameters in the MDP. In these settings, people's judgments about the goals and beliefs of agents often align closely with the inverse planning model, suggesting that our *cognitivist representations of others are undergirded by the assumption that people will tend to plan their actions roughly optimally*; deviations from optimality license inferences such as that the agent has a mistaken belief or is pursuing an alternative goal (Baker et al., 2017).

However, adults routinely make inferences not only about goals and beliefs, but also about others' *utilities*, i.e., the costs and rewards they incur when pursuing their goals. In this vein, a similar approach to the inverse planning model has also been used to account for inferences about others' *preferences* from their actions (Jern et al., 2017; Lucas et al., 2014). Here, the model is based on the assumption that others will choose among options in ways that maximize utility. Features of their utility function (e.g., preferences) can then be inferred based on their choices through Bayesian *inverse decision-making*. Collectively, these results, along with more recent work which combines inverse planning and inverse decision-making have come to be known as the *naïve utility calculus* model of intuitive psychology (Jara-Ettinger, 2019; Jara-Ettinger et al., 2017; Jara-Ettinger et al., 2016; Jara-Ettinger et al., 2020). The theoretical thrust of this account is that people's intuitive psychology is built around an expectation that those around them will act to maximize expected utility. That is, they will set goals, choose among alternatives, and even choose what *not* to do using a process which approximates rational planning. The job of the observer trying to make sense of their actions is to infer the most likely beliefs, goals, intentions, and subjective costs and rewards which would make their actions utility maximizing (Jara-Ettinger et al., 2020).

In practice, experimental tests of this theory probe people's social inferences in relatively simple grid world paradigms where agents choose among various goals to balance cost and reward. Nonetheless, the naïve utility calculus model has been shown to qualitatively capture a wide range of everyday social inferences (Jara-Ettinger et al., 2016) and to offer a precise account of more fine-grained judgments about the causes of others' behavior (Baker et al., 2017; Jara-Ettinger et al., 2020). Further, it is supported by a large body of developmental research suggesting that children are sensitive to cost, reward, and other atomic components of this account beginning in early infancy (Jara-Ettinger et al., 2017; Jara-Ettinger et al., 2016; Jara-Ettinger, Gweon, et al., 2015; Jara-Ettinger, Tenenbaum, et al., 2015; Liu & Spelke, 2017; Liu et al., 2017); this work offers a plausible explanation of how such a rich inferential paradigm in adults gets off the ground. Indeed, the strength of this account lies in part in its flexibility; while the full model captures a wide range of intuitive inferences and predictions about others' behavior even in simple settings, ablated versions of the model which represent simpler, *heuristic* decision processes often fail to fully capture the range of responses, and systematic errors, that participants make (Baker et al., 2017; Jara-Ettinger et al., 2020). This suggests that, in the settings in which the naïve utility calculus model is tested, *behaviorist* mental models based on static heuristics about the causes of others' behavior may be insufficient to capture the range of social inferences people are capable of.

Beyond its usefulness in describing people's *predictions* and *inferences* about others, having a richly structured internal model of those around us may be critical for many forms of *interaction*. For example, M. Ho et al. (2022) argue that a cognitivist theory of mind offers a unified solution to predicting others and *planning one's own* actions which will have desired outcomes involving others. By viewing others' actions as determined by goals and beliefs, we can not only predict their behavior, but select actions that will *change* their behavior, e.g., by trying to modify their goals. In addition, a large body of recent work has explored people's representation of others' mental states in *social learning* contexts (for a recent review, see Gweon (2021)). This work proposes that our ability to learn from others rests on a rich internal representation of how a teacher's intentions and beliefs guide their behavior, and further, that teachers rely on similarly sophisticated mental models of learners when choosing what information to convey (Aboody et al., 2023; Gweon, 2021; Shafto et al., 2014; Vélez & Gweon, 2019, 2021). In line with this, children's social learning appears to be supported by a representation of adults as *intentional agents* even as infants (Csibra & Gergely, 2009). Thus, in broad strokes, a central narrative in recent work exploring how we represent those around us is that *social reasoning is founded on rich cognitivist internal models of others*: we predict and interpret their behavior as if they are optimal planners given their own beliefs and knowledge, and we choose interventions and pedagogical demonstrations based on inferences about their underlying knowledge and goals.

0.5 Use cases for behaviorist intuitive psychology

While *cognitivist* models of intuitive psychology such as the naïve utility calculus provide a flexible account of how we represent others based on their behavior, the complexity of these models poses a number of challenges that might in principle be resolved by simpler *behaviorist* mental models. Here, I review some of these challenges and what alternatives, if any, have been proposed to address them. Broadly, these concerns fall into two classes: settings in which observers may fail to reason accurately about others' mental states (even if they should), and settings in which people's behavior—or the best explanation for it—falls outside the domain of cognitivist mental models.

0.5.1 Failures of cognitivist intuitive psychology

First, the expectation that others are rational (or even approximately rational) poses computational challenges for the *observer*, even if it is the "right" model. Experimental validation of the naïve utility calculus account has relied primarily on grid world settings where individual agents can be modeled using Markov Decision Processes (Bellman, 1957). The observer then makes mental state inferences by performing *inverse reinforcement learning* (IRL) over the actions of others (Jara-Ettinger, 2019). In spite of recent gains in computational power to support these models, IRL remains a computationally challenging process, especially in domains outside the simplified grid world paradigm (Jara-Ettinger, 2019). While it's been argued that *approximate* solutions may be both tractable and sufficient for modeling human inferences in other settings (Baker et al., 2017), this remains an open challenge.

Consistent with the computational overhead imposed by reasoning about others as optimal planners, experimental work has shown that in less *deliberative* settings, people find theory of mind reasoning effortful and may forego it altogether (Epley et al., 2004; Keysar et al., 2000; Keysar et al., 2003; Lin et al., 2010). For example, in one such study (Keysar et al., 2003), participants were paired with a confederate for a reference game in which the confederate would give the participant repeated instructions to move different objects between slots in a vertical grid placed between them. Before the experiment began (and unbeknownst to the confederate), participants were instructed to hide a particular item (e.g., a roll of tape) in a paper bag and place it in one of several slots in the grid that were occluded from the confederate. Then, during *critical trials* throughout the experiment, the confederate would instruct participants to move a mutually visible item using instructions that were ambiguous when also considering the hidden item, but not ambiguous with respect to what the *confederate knew* about the grid. For example, the confederate might instruct the participant to move a cassette tape that was visible to both, saying simply, "move the tape" (since the confederate was unaware of the hidden roll of tape, this could only refer to the cassette). Across the eight critical trials, 71% of participants attempted to move the hidden item at least once and 46% did so in at least half of the trials (Keysar et al., 2003). Such behavior seems to reflect a failure to incorporate the confederate's beliefs into participants' own decisions, despite having full awareness that the confederate *can* see one of the items and is unaware of the ambiguous second item. The authors argue that in situations where people have less time or resources for deliberative processing, people are "mindblind" (Lin et al., 2010), failing to exhibit theory of mind reasoning about others' beliefs that they are fully capable of otherwise.

However, a failure to optimally integrate others' beliefs into one's own behavior has been shown in more deliberative contexts as well; for example, recent work by Aboody et al. (2023) finds that in pedagogical settings where participants were instructed to choose informative examples to illustrate how a machine worked, participants failed to account for the full set of beliefs that learners held about the machine. Thus, research on *failures* of traditional cognitivist intuitive psychology suggests that in domains where the space of mental states is vast (e.g., beliefs about how a machine works) or the task itself admits less opportunity for deliberation, people may struggle to infer others' beliefs or integrate these beliefs into their own actions. However, it remains unclear to what extent, if at all, people employ a simpler *behaviorist* internal model of others in these settings.

0.5.2 Predicting habits and other patterned behaviors

At a high-level, research on people's representation of others has largely focused on settings where inferring others' mental states such as goals or beliefs is useful *by design*. But what about settings where this is not the case, where people's actions are best explained by simpler heuristics, habits, or other processes? One concern frequently addressed by proponents of the naïve utility calculus centers on the underlying assumption that others are rational decision-makers. One might object that this is a misguided hypothesis about our representations of others since people often fail to exhibit rational behavior in the first place. People make mistakes, they forget where they left their keys, and they display a complex and well-documented suite of biases and judgment errors when making decisions (Kahneman, 2011; Tversky & Kahneman, 1974). We know this about ourselves and those around us; if rational action is a poor fit to individual behavior, what makes it a reasonable hypothesis about how we represent *others*? Proponents of the naïve utility calculus point out that a model of others need not have perfect fidelity to be useful and that in fact, our tendency to view others this way might explain why some violations of rationality are so surprising (Baker et al., 2017; Jara-Ettinger et al., 2016). Recent work has also explored the use of *boundedly rational* generative processes within this same framework to account for scenarios where people fail to show optimal behavior due to confusion of goals, misguided planning, or errors in action execution (Alanqary et al., 2021).

However, there are well-known settings where people's behavior is not a result of rational (or even *boundedly* rational) planning; intuition suggests that our representations of others can accommodate this. Consider for example somebody acting out of *habit*. Habits are a regular and important part of our day-to-day behavior (Bargh & Chartrand, 1999) and are often modeled as actions that result from simple, learned associations with rewards in our environment (Daw et al., 2005; Dolan & Dayan, 2013; K. J. Miller et al., 2019; Wood & Neal, 2007). Thus, habitual action is likely *best* understood using a behaviorist intuitive psychology. Gershman et al. (2016) found that people's predictions about whether an agent would choose a habitual action or a non-habitual (but more optimal) action were sensitive to many of the features that underlie habitual action in the first place: repetition, consistency with other tasks, decision time, and cognitive load. In short, people's intuitive psychology of habitual behavior appears closely aligned with empirical and formal treatments of it. Critically, these results also show that people can, and do, apply a behaviorist internal model of others—in place of a cognitivist, mental-state driven model—when predicting habitual action.

Habits are not the only form of individual action that might be interpreted with a behaviorist model of others. Looked at more broadly, a great many human behaviors exhibit strong statistical regularities which could form the basis for predicting actions. In some activities, like tennis or other fast-paced sports, it seems unlikely that people can perform the requisite *cognitivist* predictions at the speed that would be needed, yet statistical structure in an opponent's shot patterns might allow for reasonable prediction. Other activities, such as cooking dinner, occur at a pace that permits mental state inferences, yet these may not be strictly necessary for predicting action; adding pasta to a pan *after* boiling the water likely has little to do with desires or beliefs *per se*, but simply the scripted nature of this behavior. To what degree do people reason about others' actions using a behaviorist mental model in such settings?

Early work on human *statistical learning* highlighted infants' ability to detect structure in transition probabilities between novel syllables (Saffran et al., 1996), as well as non-linguistic tones (Saffran et al., 1999); in adults, these findings have been extended to visual sequences as well (Fiser & Aslin, 2002; Turk-Browne et al., 2005), providing a foundation for behaviorist intuitive psychology. But do we leverage this ability for social inferences? Prior work has shown that adults and toddlers can predict the subsequent actions of individuals who are playing when these people show statistical regularities in their action sequences (Monroy et al., 2017). In fact, some have argued that infants' success on theory of mind tasks may be at least partially supported by statistical learning (Ruffman et al., 2012). Critically, predicting others based on known statistical regularities may be something we do for a broad range of everyday actions; Thornton and Tamir (2017) find that people have calibrated estimates of transition probabilities between emotional states and that this gives them the ability to predict several steps forward in large datasets of experienced emotions. Building on these results, the authors explore people's ability to predict action transitions from among a large set of naturalistic behaviors. Given prompts such as, "how likely is someone to start running, given that they are currently stretching?" they find that people are well calibrated relative to naturalistic sequences extracted from movie scripts, the American Time Use Survey, WikiHow, and other similar resources (Thornton & Tamir, 2021).

The finding that people are sensitive to the development of habit in others, and to predictive regularities underlying a host of day-to-day activities, suggests that people are capable of leveraging a behaviorist mental model in many settings where doing so is useful. When considering the range of *other* ways in which individual action might be predicted using learnable patterns rather than mental states like goals and beliefs—norms (Sarathy et al., 2017), conventions (Hawkins et al., 2019), or scripted behavior (Schank & Abelson, 1977)—it may be that behaviorist mental models play an even broader role in our social reasoning. They may also serve to reduce the computational costs of inferences about others when full-fledged cognitivist mental models are overly cumbersome. However, as the results above suggest, existing literature lacks a coherent framework for how and when people rely on a behaviorist intuitive psychology, what the limits of this ability are, and how intelligent reasoners might trade off between more *cognitivist* and *behaviorist* representations. The work presented here attempts to make strides in this direction.

0.6 Current work

The research presented in this dissertation takes up the broad question discussed at the outset: *what is the structure of our mental representations of others?* I argue that we can arrive at a useful conceptual framing of this question by appeal to existing work in psychology on how individuals make decisions—our representations of others should at some level corroborate our best understanding of the ways people themselves act. To this end, I describe two models of individual behavior that have predominated cognitive psychology and artificial intelligence for the last 100 years: *behaviorism* and *cognitivism.* While the way people infer mental states like beliefs and goals—the basis for their *cognitivist* representations—has received considerable attention, spurred by recent advances in computational cognitive modeling, we lack a coherent account of how and when people might rely on more *behaviorist* models of others. Yet existing work provides evidence that this is both possible and useful for everyday prediction (Gershman et al., 2016; Thornton & Tamir, 2021).

What might a good test of pattern-based, behaviorist mental models look like? In Chapter 1, I propose that repeated interactions using *mixed strategy equilibrium* (MSE) games like rock, paper, scissors presents an ideal use case for understanding how people build up mental models of an opponent based only on patterns in their prior actions. This approach differs from the research described above in several important ways. First, unlike statistical regularities in everyday action sequences, patterns in repeated games must be detected from scratch. Second, unlike experimental settings used in prior theory of mind work, in which people might plausibly use a combination of cognitivist and behaviorist representations, the *only* way to succeed over many rounds of interaction in MSE games is to build up a predictive model based on regularities in the opponent's moves. Finally, an important contribution of this work is to show how the patterns that might form the basis for people's opponent modeling can be clearly spelled out at increasing levels of complexity. Thus, while prior work has largely ignored the question of how complex our behaviorist mental models *can be*, results in chapter 1 illustrate how MSE games provide a tractable way to approach this question. This work was originally published in Brockbank and Vul (2021) and is re-printed here with minor edits.

Empirical results analyzed in chapter 1 suggest that people do indeed build up predictive internal models of their opponents over many rounds of interaction in MSE games. This raises questions about the underlying *content* of people's behaviorist mental models in this setting. What kinds of sequential patterns are people capable of adapting to in their opponent's moves? And what kinds of patterns can they revise in their own? In chapter 2, I address these questions with results from two studies where people play many rounds of rock, paper, scissors against an algorithmic opponent. These opponents display a range of increasingly complex patterns in their move choices (experiment 1) and try to exploit a range of patterns in participant move choices (experiment 2). By studying participants' success against these opponents, we are able to map out the space of sequential patterns that people can adapt to and reduce in their own moves. In this way, we provide an initial answer to the *underlying structure* of people's behaviorist mental models. The work in this chapter has been submitted for publication and is currently under revision (Brockbank & Vul, 2023).

Chapters 1 and 2 focus on people's ability to learn increasingly complex representations of others' sequential behavior. In chapter 3, I ask how people might build up the sophisticated mental models of others typical of more cognitivist representations. In this chapter, I also seek to widen the scope of my inquiry by focusing on cooperative interactions in a physical inference task where people have a high degree of domain knowledge. This chapter begins by asking how people learn single latent parameters which guide a confederate's behavior (in this case, their accuracy); I focus on people's ability to *dynamically update* these estimates with collaborators that improve and worsen over time, and how much they show evidence of *relying* on this latent parameter estimate to guide collaborative behavior. In a second study, I explore people's ability to learn information from their partner's behavior which is diagnostic of their partner's *underlying model of the task*. In this way, my co-authors and I explore the limits of people's ability to develop richly structured, task-based models of another agent. Study 1 was previously published in an earlier form (Brockbank et al., 2022); this chapter includes new work and is currently being prepared for submission.

Taken together, the work in these three chapters shows that by studying repeated, dyadic interactions in settings which allow for complex behavior and sophisticated mental models of others, we achieve insights about the *structure* of people's representations of those around them. In the final chapter, I conclude with a brief discussion of what these results tell us, and how they might inform future work aimed at addressing several key questions: a) how people model those around them at the right level of complexity for a given situation, b) how people form more stable, abstract representations of others, and c) how people transfer their representations of others to novel tasks or contexts. The work in this dissertation offers novel experimental and computational approaches to addressing these questions and a set of results which support deeper inquiry into our remarkable ability to represent other humans in rich and flexible ways.

References

Aboody, R., Velez-Ginorio, J., Santos, L. R., & Jara-Ettinger, J. (2023). When naïve pedagogy breaks down: Adults rationally decide how to teach, but misrepresent learners' beliefs. *Cognitive Science*, *47*(3), e13257.

Alanqary, A., Lin, G. Z., Le, J., Zhi-Xuan, T., Mansinghka, V. K., & Tenenbaum, J. B. (2021). Modeling the mistakes of boundedly rational agents within a bayesian theory of mind. *arXiv preprint arXiv:2106.13249*.

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), 0064.

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition*, *113*(3), 329–349.

Bargh, J. A., & Chartrand, T. L. (1999). The unbearable automaticity of being. *American psychologist*, *54*(7), 462.

Bellman, R. (1957). A markovian decision process. *Journal of mathematics and mechanics*, 679–684.

Bigelow, E., & Ullman, T. D. (2022). People's evaluation of programs that drive agents' behavior. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *44*(44).

Brockbank, E., & Vul, E. (2023). Rock, paper, scissors play reveals limits in adaptive sequential behavior. *Manuscript submitted for publication.*

Brockbank, E., & Vul, E. (2021). Humans fail to outwit adaptive rock, paper, scissors opponents. In T. Fitch, C. Lamm, H. Leder, & K. Teßmar-Raible (Eds.), *Proceedings of the 43rd annual conference of the cognitive science society* (pp. 1740–1746). Cognitive Science Society.

Brockbank, E., *Wang, H., Yang, J., Mirchandani, S., Bıyık, E., Sadigh, D., & Fan, J. E. (2022). How do people incorporate advice from artificial agents when making physical judgments? In J. Culbertson, H. Rabagliati, V. Ramenzoni, & A. Perfors (Eds.), *Proceedings of the 44th annual conference of the cognitive science society* (pp. 1375–1381). Cognitive Science Society.

Callaway, F., van Opheusden, B., Gul, S., Das, P., Krueger, P. M., Griffiths, T. L., & Lieder, F. (2022). Rational use of cognitive resources in human planning. *Nature Human Behaviour*, *6*(8), 1112–1125.

Chomsky, N. (1957). *Syntactic structures*. Mouton & Co.

Csibra, G., & Gergely, G. (2009). Natural pedagogy. *Trends in cognitive sciences*, *13*(4), 148–153.

Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature neuroscience*, *8*(12), 1704–1711.

Dayan, P., & Niv, Y. (2008). Reinforcement learning: The good, the bad and the ugly. *Current opinion in neurobiology*, *18*(2), 185–196.

Dolan, R. J., & Dayan, P. (2013). Goals and habits in the brain. *Neuron*, *80*(2), 312–325.

Epley, N., Keysar, B., Van Boven, L., & Gilovich, T. (2004). Perspective taking as egocentric anchoring and adjustment. *Journal of personality and social psychology*, *87*(3), 327.

Fiser, J., & Aslin, R. N. (2002). Statistical learning of higher-order temporal structure from visual shape sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(3), 458.

Gardner, H. (1987). *The mind's new science: A history of the cognitive revolution*. Basic books.

Gershman, S. J., Gerstenberg, T., Baker, C. L., & Cushman, F. A. (2016). Plans, habits, and theory of mind. *PloS one*, *11*(9), e0162246.

Goodman, N. D., Baker, C. L., Bonawitz, E. B., Mansinghka, V. K., Gopnik, A., Wellman, H., Schulz, L., & Tenenbaum, J. B. (2006). Intuitive theories of mind: A rational approach to false belief. *Proceedings of the twenty-eighth annual conference of the cognitive science society*, *6*.

Gopnik, A., & Wellman, H. M. (1992). Why the child's theory of mind really is a theory.

Gweon, H. (2021). Inferential social learning: Cognitive foundations of human social learning and teaching. *Trends in Cognitive Sciences*, *25*(10), 896–910.

Hawkins, R. X., Goodman, N. D., & Goldstone, R. L. (2019). The emergence of social norms and conventions. *Trends in cognitive sciences*, *23*(2), 158–169.

Heider, F. (1958). *The psychology of interpersonal relations*. Lawrence Erlbaum Associates.

Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *The American journal of psychology*, *57*(2), 243–259.

Ho, M., Saxe, R., & Cushman, F. (2022). Planning with theory of mind. *Trends in Cognitive Sciences*.

Ho, M. K., Abel, D., Correa, C. G., Littman, M. L., Cohen, J. D., & Griffiths, T. L. (2022). People construct simplified mental representations to plan. *Nature*, *606*(7912), 129–136.

Jara-Ettinger, J. (2019). Theory of mind as inverse reinforcement learning. *Current Opinion in Behavioral Sciences*, *29*, 105–110.

Jara-Ettinger, J., Floyd, S., Tenenbaum, J. B., & Schulz, L. E. (2017). Children understand that agents maximize expected utilities. *Journal of Experimental Psychology: General*, *146*(11), 1574.

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.

Jara-Ettinger, J., Gweon, H., Tenenbaum, J. B., & Schulz, L. E. (2015). Children's understanding of the costs and rewards underlying rational action. *Cognition*, *140*, 14–23.

Jara-Ettinger, J., Schulz, L. E., & Tenenbaum, J. B. (2020). The naive utility calculus as a unified, quantitative framework for action understanding. *Cognitive Psychology*, *123*, 101334.

Jara-Ettinger, J., Sun, F., Schulz, L., & Tenenbaum, J. B. (2018). Sensitivity to the sampling process emerges from the principle of efficiency. *Cognitive Science*, *42*, 270–286.

Jara-Ettinger, J., Tenenbaum, J. B., & Schulz, L. E. (2015). Not so innocent: Toddlers' inferences about costs and culpability. *Psychological science*, *26*(5), 633–640.

Jern, A., Lucas, C. G., & Kemp, C. (2017). People learn other people's preferences through inverse decision-making. *Cognition*, *168*, 46–64.

Kahneman, D. (2011). *Thinking, fast and slow*. macmillan.

Keysar, B., Barr, D. J., Balin, J. A., & Brauner, J. S. (2000). Taking perspective in conversation: The role of mutual knowledge in comprehension. *Psychological Science*, *11*(1), 32–38.

Keysar, B., Lin, S., & Barr, D. J. (2003). Limits on theory of mind use in adults. *Cognition*, *89*(1), 25–41.

Lin, S., Keysar, B., & Epley, N. (2010). Reflexively mindblind: Using theory of mind to interpret behavior requires effortful attention. *Journal of Experimental Social Psychology*, *46*(3), 551–556.

Liu, S., & Spelke, E. S. (2017). Six-month-old infants expect agents to minimize the cost of their actions. *Cognition*, *160*, 35–42.

Liu, S., Ullman, T. D., Tenenbaum, J. B., & Spelke, E. S. (2017). Ten-month-old infants infer the value of goals from the costs of actions. *Science*, *358*(6366), 1038–1041.

Lucas, C. G., Griffiths, T. L., Xu, F., Fawcett, C., Gopnik, A., Kushnir, T., Markson, L., & Hu, J. (2014). The child as econometrician: A rational model of preference understanding in children. *PloS one*, *9*(3), e92160.

Mandler, G. (2002). Origins of the cognitive (r) evolution. *Journal of the History of the Behavioral Sciences*, *38*(4), 339–353.

Miller, G. A. (2003). The cognitive revolution: A historical perspective. *Trends in cognitive sciences*, *7*(3), 141–144.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review*, *63*(2), 81.

Miller, K. J., Shenhav, A., & Ludvig, E. A. (2019). Habits without values. *Psychological review*, *126*(2), 292.

Monroy, C., Meyer, M., Gerson, S., & Hunnius, S. (2017). Statistical learning in social action contexts. *PloS one*, *12*(5), e0177261.

Newell, A., Shaw, J. C., & Simon, H. A. (1958). Elements of a theory of human problem solving. *Psychological review*, *65*(3), 151.

Núñez, R., Allen, M., Gao, R., Miller Rigoli, C., Relaford-Doyle, J., & Semenuks, A. (2019). What happened to cognitive science? *Nature human behaviour*, *3*(8), 782–791.

Onishi, K. H., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *science*, *308*(5719), 255–258.

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and brain sciences*, *1*(4), 515–526.

Ruffman, T., Taumoepeau, M., & Perkins, C. (2012). Statistical learning as a basis for social understanding in children. *British Journal of Developmental Psychology*, *30*(1), 87–104.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926–1928.

Saffran, J. R., Johnson, E. K., Aslin, R. N., & Newport, E. L. (1999). Statistical learning of tone sequences by human infants and adults. *Cognition*, *70*(1), 27–52.

Sarathy, V., Scheutz, M., & Malle, B. F. (2017). Learning behavioral norms in uncertain and changing contexts. *2017 8th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)*, 000301–000306.

Schank, R. C., & Abelson, R. P. (1977). *Scripts, plans, goals, and understanding: An inquiry into human knowledge structures*. Routledge.

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.

Skinner, B. F. (1938). *The behavior of organisms*. Appleton-Century-Crofts.

Sutton, R. S., & Barto, A. G. (1998). *Introduction to reinforcement learning* (Vol. 135). MIT press Cambridge.

Thorndike, E. L. (1911). *Animal intelligence: Experimental studies*. Macmillan.

Thornton, M. A., & Tamir, D. I. (2017). Mental models accurately predict emotion transitions. *Proceedings of the National Academy of Sciences*, *114*(23), 5982–5987.

Thornton, M. A., & Tamir, D. I. (2021). People accurately predict the transition probabilities between actions. *Science Advances*, *7*(9), eabd4995.

Turk-Browne, N. B., Jungé, J. A., & Scholl, B. J. (2005). The automaticity of visual statistical learning. *Journal of Experimental Psychology: General*, *134*(4), 552.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *science*, *185*(4157), 1124–1131.

Ullman, T., Baker, C., Macindoe, O., Evans, O., Goodman, N., & Tenenbaum, J. (2009). Help or hinder: Bayesian models of social goal inference. *Advances in neural information processing systems*, *22*.

Vélez, N., & Gweon, H. (2019). Integrating incomplete information with imperfect advice. *Topics in cognitive science*, *11*(2), 299–315.

Vélez, N., & Gweon, H. (2021). Learning from other minds: An optimistic critique of reinforcement learning models of social learning. *Current Opinion in Behavioral Sciences*, *38*, 110–115.

Watson, J. B. (1924). *Behaviorism*. Routledge.

Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development: The truth about false belief. *Child development*, *72*(3), 655–684.

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, *13*(1), 103–128.

Wood, W., & Neal, D. T. (2007). A new look at habits and the habit-goal interface. *Psychological review*, *114*(4), 843.

Zhi-Xuan, T., Mann, J., Silver, T., Tenenbaum, J., & Mansinghka, V. (2020). Online bayesian goal inference for boundedly rational planning agents. *Advances in neural information processing systems*, *33*, 19238–19250.

Chapter 1

Formalizing opponent modeling with the rock, paper, scissors game

Abstract

In simple dyadic games such as rock, paper, scissors (RPS), people exhibit peculiar sequential dependencies across repeated interactions with a stable opponent. These regularities seem to arise from a mutually adversarial process of trying to outwit their opponent. What underlies this process, and what are its limits? Here, we offer a novel framework for formally describing and quantifying human adversarial reasoning in the rock, paper, scissors game. We first show that this framework enables a precise characterization of the complexity of patterned behaviors that people exhibit themselves, and appear to exploit in others. This combination allows for a quantitative understanding of human opponent modeling abilities. We apply these tools to an experiment in which people played 300 rounds of RPS in stable dyads. We find that although people exhibit very complex move dependencies, they cannot exploit these dependencies in their opponents, indicating a fundamental limitation in people's capacity for adversarial reasoning. Taken together, the results presented here show how the rock, paper, scissors game allows for precise formalization of human adaptive reasoning abilities.

Keywords: adversarial reasoning, sequential reasoning, competition, rock-paper-scissors

1.1 Introduction

At a basic level, human conflict and coordination is rooted in the ability to predict the behavior of others and make plans accordingly. While this may sometimes involve ad-hoc coordination from first principles, such as well-known Schelling point behavior (Schelling, 1960), more often we find ourselves in repeated interactions, wherein we have the opportunity to adapt to past outcomes. Everyday life is replete with such dynamics, whether playing basketball or chess, or simply commuting in traffic among other drivers that are all trying to get home as fast as possible. Broadly, competitive interactions highlight our ability to anticipate and respond to others in diverse settings. What cognitive processes underlie our remarkable ability to anticipate and adapt to the behavior of others around us across repeated interactions? We argue that this question can be addressed by examining people's behavior in repeated adversarial games, such as rock-paper-scissors, where success is a matter of outsmarting one's opponent, often by identifying predictable patterns in their choices.

To better understand how people manage the cognitive challenges of adapting to others in adversarial interactions, researchers have traditionally turned to iterated zero-sum games. Zero-sum games have the unique character that any player's gain comes at a loss to their opponent: they are the "limiting case of pure conflict" (Schelling, 1958). Here, we focus on the game of rock, paper, scissors (RPS), or roshambo. In this game, two players simultaneously produce a hand signal indicating their choice of "rock", "paper", or "scissors". The rules are simple: "rock" beats "scissors", "paper" beats "rock", and "scissors" beats "paper". The game is perhaps most popular with children, but it has been used in official contexts to settle court disputes (Liptak, 2006) and art auctions (Vogel, 2005). Large scale RPS tournaments have been held with human entrants (Hegan, 2004), while the potential to test a diverse set of algorithmic strategies has also inspired tournaments modeled after Axelrod (1984) in which various bots compete against each other (Billings, 2000a, 2000b) (and more recently on the data science site Kaggle, see: https://www.kaggle.com/c/rock-paper-scissors). Finally, the dynamics of the game have made it popular for modeling diverse biological ecosystems (Allesina & Levine, 2011; Claussen & Traulsen, 2008; Kerr et al., 2002; Kirkup & Riley, 2004; Sinervo & Lively, 1996; Zhang et al., 2013), offering predictions in evolutionary game theory (Garrido-da-Silva & Castro, 2020; Hu et al., 2019; Szolnoki et al., 2014; Toupo & Strogatz, 2015; Yang et al., 2017), and even studying large scale market behavior (Cason et al., 2005; Hauert et al., 2002; Hopkins & Seymour, 2002; Lach, 2002; Lakhar, 2011; Noel, 2007; Semmann et al., 2003).

Beyond its role in popular culture and in various academic disciplines, the rock, paper, scissors game offers a unique means of studying human adversarial behavior during repeated interactions. Here, our focus is on decision making across many iterated rounds against a stable opponent—often hundreds, rather than the "best of 3" used to resolve household disputes. In such laboratory studies of rock, paper, scissors, the large number of interactions allow people to detect and adapt to potentially complex patterns in their opponent's behavior. In fact, due to the game's simple rules and constrained space of choices, better performance by one individual over many rounds will not likely be a result of general game "expertise", but rather a result of superior reasoning about dependencies in their specific opponent's move choices. This reliance on adaptation to a particular opponent, rather than general game expertise, distinguishes RPS from other adversarial games like chess, and makes it a purer form of adversarial reasoning. Finally, RPS, like other mixed strategy equilibrium games, is characterized by its Nash Equilibrium solution (Nash, 1950) which dictates random move selection, a strategy which presents unique cognitive challenges for human players. For these reasons, a large body of literature has examined human behavior over repeated interactions in the rock, paper, scissors game, motivated by diverse questions about the nature of human learning, sequential behavior, and perceptions of randomness (Budescu & Rapoport, 1994; Dyson, 2019; Zhou, 2016).

In the present work, we argue that the rock, paper, scissors game represents an ideal means of studying human *adaptive, adversarial* reasoning capacities, i.e., the ability to outwit another person by discovering patterns in their behavior, and offer a novel set of results illustrating the limits of this ability. First, we briefly examine the findings from previous literature on the rock, paper, scissors game with an eye to what existing results tell us about human adversarial reasoning. We argue that by focusing on failures of Nash Equilibrium and on coarse heuristics, prior work has largely overlooked the question of how people adapt to a fallible human opponent over repeated interactions. In this vein, we next discuss how the structure of the game offers a tractable way of describing the flexibility and limitations of people's adaptive reasoning capacities. To illustrate this, we present an analysis of existing results which suggests that the ability to recognize and exploit sequential patterns in RPS is highly constrained, revealing the limits of human adaptive reasoning.

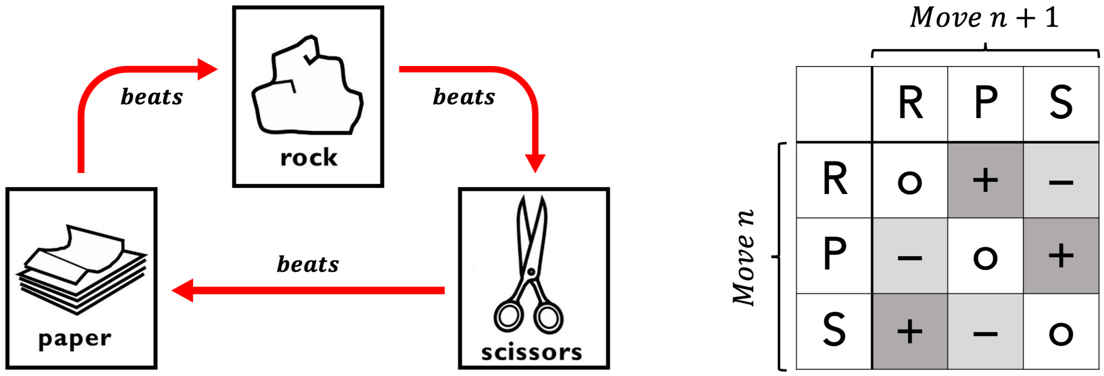
1.2 Human RPS behavior reflects adversarial reasoning

First, we consider what is known about human behavior in iterated rock, paper, scissors games. This literature often starts with the behavioral economics perspective of comparing human behavior to optimal play and, upon finding a difference, seeks to explain it in terms of human heuristics or biases. In RPS, optimal behavior is taken to be uniform random choices, and failures to achieve such randomness are explained as human failures to generate random sequences. Here we instead argue that the deviations from optimality documented in this literature are more consistent with people attempting to adapt to, and outwit, their opponent, rather than trying and failing to generate truly random move choices. In short, we argue that the existing literature supports the claim that human RPS behavior reflects adaptive adversarial reasoning.

1.2.1 Normative strategies

The starting point for exploring human behavior in the rock, paper, scissors game has traditionally focused on whether people adhere to the normative standards of Nash Equilibrium (Nash, 1950), in which a strategy is chosen to optimize performance under the assumption of an equivalently rational, optimizing opponent. RPS belongs to the class of zero-sum *cyclic dominance* games (Morgenstern & Neumann, 1953). Their cyclic nature is best illustrated with the well-known rules of RPS, where "rock" beats "scissors" and "paper" beats "rock", but "paper" is beaten by "scissors" (see Figure 1.1 for an illustration of this). Thus, every choice is dominated by one other and no choice is better than another, unless you have some information about what the opponent will choose. Such games are not limited to three-choice paradigms like RPS; cyclic games with many more choices provide a unique means of studying large-scale group behaviors (Frey & Goldstone, 2013).

Given that no move is better than any other in a cyclic dominance game, how *should* one make strategic decisions in rock, paper, scissors? The zero-sum nature of the game ensures that for a single player, their opponent's win is always their loss, so any degree to which a player's decisions are predictable will allow their opponent to exploit them for a greater gain. Therefore, the best strategy for a rational player paired with an equally rational opponent is to choose moves so as to not create any exploitable patterns in their choices: to choose the three options randomly, with equal probability. Cyclic dominance games belong to the broader class of mixed strategy equilibrium (MSE) games (see Camerer (2011) ch. 3 for review), with a single Nash Equilibrium (NE) (Nash, 1950) that requires a mixed strategy of playing each move (e.g., "rock", "paper", and "scissors") in equal proportion, with no conditional dependence from one game to the next. Indeed, the appeal of studying decision making in RPS and other similar games has been in large part due to the fact that they impose such strong, testable constraints on optimal play; constraints that human behavior often fails to exhibit.



	R	P	S
R	o	+	–
P	–	o	+
S	+	–	o

(a) Cyclic dominance in rock, paper, scissors

(b) Categorizing move transitions

Figure 1.1. The rock, paper, scissors game. (a) Shows the cyclic dominance relations of the three move choices: "rock" beats "scissors", "scissors" beats "paper", "paper" beats "rock". (b) These cyclic dominance relations mean that the relationship between one move and the next can be characterized into one of three "transitions": a "positive" transition or shift "up" to the move that would beat the previous move (+), a "negative" transition or shift "down" to the move that would lose to the previous move (−), and a "stay" transition which repeats the same move (0).

1.2.2 Human behavior exhibits sequential patterns

Some of the earliest research in mixed strategy equilibrium games like RPS puzzled over whether people could in fact meet the high standards of random play under the Nash Equilibrium strategy (Brown & Rosenthal, 1990; Kalisch et al., 1954; O'Neill, 1987); for an overview of significant early results, see Camerer (2011) ch. 3. A large body of work has shown that in rock, paper, scissors and other MSE games, people exhibit a range of sequential regularities or *dependencies* in their move choices that run counter to equilibrium play. A full review of these results is beyond the scope of the current paper, but here we offer a sample, surveying evidence for sequential dependencies in order of increasing behavioral complexity (Dyson, 2019).

A first pass analysis of people's behavior in the rock, paper, scissors game often looks at whether their *overall* distribution of move choices is consistent with the mixed strategy equilibrium proportions of 1/3 for each move. In repeated games of RPS, a number of studies have found people to have a slight overall bias towards “rock”, though this is not always significant (Dyson et al., 2020; Dyson et al., 2016; Forder & Dyson, 2016; Wang et al., 2014; Xu et al., 2013). Further, other results have observed a modest preference for “paper” or “scissors” (Aczel et al., 2012) and in many cases people show no distinguishable preference at all (Cook et al., 2012; Kangas et al., 2009; Lie et al., 2013; Stöttinger, Filipowicz, Danckert, et al., 2014). In the broader space of MSE games, Camerer (2011) notes that marginal choice probabilities tend to align with equilibrium proportions.

Though marginal move distributions are often approximately consistent with equilibrium random selection, a key feature of the Nash Equilibrium strategy is that players not display any conditional dependence on their own or their opponents’ previous moves. Thus, a player that continually cycles from “rock” to “paper” to “scissors” will produce an overall distribution of moves that appears identical to the mixed strategy equilibrium but the statistical dependence on their own previous move will be highly exploitable by a perceptive opponent. Following prior work (Dyson, 2019), we will refer to a transition from one move to the move that beats it (e.g., “rock” to “paper”) as *shifting up* (denoted with a + in tables and figures); a transition from one move to the same move (e.g., “rock” to “rock”) as *staying* (denoted with a 0 in tables and figures), and a transition from one move to the move that loses to it (e.g., “rock” to “scissors”) as *shifting down* (denoted with a $-$ in tables and figures). See Figure 1.1 for a complete illustration of the transitions between moves.

Evidence of transition dependencies in people’s moves is not widespread, but Wang et al. (2014) find a slight overall preference for staying compared to shifting up or down which diminishes with the relative value of wins over ties, suggesting that stronger reward incentives may improve people’s tendency to approximate equilibrium play. Indeed, related work has argued for a relationship between transition dependencies in competitive settings and limitations in executive control; Baek et al. (2013) found that people with schizophrenia had a strong dependence on their *opponent's* previous move, tending to select moves that would beat what their opponent had just played (this is often referred to as a Cournot Best Response strategy (Cournot, 1838)). Finally, Dyson et al. (2016) find evidence for a *stickiness* of transition dependencies, namely that participants who shifted up in a previous transition were more likely to continue shifting up and participants who shifted down in a previous transition were more likely to shift down again (no such persistence was found for staying).

The best documented higher-order move dependencies in rock, paper, scissors are *transitions conditioned on prior outcome*. This is exemplified by *win-stay, lose-shift* (WSLS) behavior. In the context of rock, paper, scissors, such a strategy amounts to changing the rates of particular transitions (+, −, 0) depending on whether the preceding game outcome was a win, loss, or tie. The appeal of WSLS as a possible explanatory mechanism for people's decisions in games like RPS comes from its prominence in other settings where it can be seen as a computationally simple heuristic that enables broadly adaptive behavior (Gigerenzer & Goldstein, 1996; Posch, 1999). A number of studies have found evidence of outcome-dependent transition behavior in rock, paper, scissors (Cason et al., 2014; Hoffman et al., 2012; Wang & Xu, 2014; Wang et al., 2014; Xu et al., 2013). Subsequent work has further explored the separability of win-stay and lose-shift behaviors (Forder & Dyson, 2016), as well as the factors mediating their respective magnitudes (Dyson et al., 2020; Dyson et al., 2018; Dyson et al., 2016).

Finally, Brockbank and Vul (2020) find that in many rounds of paired human dyad play, people exhibit a range of additional dependencies, with more complex dependencies being more pronounced. Taken together, these results have broad agreement that people's move choices exhibit unique sequential dependencies which violate NE. This raises an important question: given the failure to implement equilibrium strategies in mixed strategy games like RPS, what accounts for people's behavior, particularly the various sequential dependencies in their move choices?

1.2.3 Existing accounts of empirical behavior are insufficient

The most prominent account of why human behavior in rock, paper, scissors and other MSE games displays such sequential dependencies focuses on people's misapprehensions about what it means to be random in the first place. A large body of work on *subjective randomness* has revealed that people often have poor intuitions about what constitutes a random sequence (Bar-Hillel & Wagenaar, 1991; Lopes, 1982). Concretely, when prompted to evaluate or produce a sequence of simulated coin flips (or simulate any other random variable) people tend to favor sequences that (i) have an equal number of heads and tails, (ii) under-represent "runs" (e.g., HHH) and (iii) over-represent alternations (HTH) (Lopes & Oden, 1987; Tversky & Kahneman, 1971). In a series of studies exploring these biases in adversarial settings, Rapoport and Budescu propose a model in which randomness is a matter of "local representativeness" across a limited memory of prior events (Budescu & Rapoport, 1994; Rapoport & Budescu, 1992, 1997). Essentially, their model suggests that behavior in mixed strategy equilibrium games like rock, paper, scissors represents people doing their best to produce random outcomes. With only a limited memory for prior events, participants will make choices that exemplify the features of *subjective randomness* exhibited in prior literature.

While there is ample evidence that our judgments of random events depart systematically from true randomness, this is unlikely to explain human behavior in repeated games of rock, paper, scissors. Empirical support for behaviors that show a conditional dependence on *opponent choices* and *prior outcomes* suggests that people are doing something more complicated than merely attending to the (subjective) randomness of their own move choices (see West and Lebiere (2001) for discussion of complex opponent-responsive properties). What then can explain people's behavior, particularly the sequential patterns they exhibit, in repeated MSE games?

Another common explanation is that people may be using stable heuristics that produce winning, or at least adequate, outcomes in the long run. For instance, win-stay, lose-shift (WSLS) is a "fast and frugal" decision rule (Gigerenzer & Goldstein, 1996) that can be applied in a variety of adversarial settings; indeed, WSLS outperforms the well-known "tit-for-tat" strategy in evolutionary Prisoner's Dilemma simulations (M. A. Nowak & Sigmund, 1992). This finding fits within a broad literature on the evolution of cooperation examining the strength of various heuristic-based strategies across many interactions, though such findings typically describe population dynamics rather than individual behavior (Axelrod, 1984; M. Nowak & Sigmund, 1993; M. Nowak & Sigmund, 1990; M. A. Nowak & Sigmund, 2004). Nonetheless, fixed heuristics like WSLS might drive people's choices in repeated adversarial interactions and may explain behavioral regularities in the rock, paper, scissors game (Dyson et al., 2016; Wang et al., 2014; Zhou, 2016). Dyson et al. (2018) propose a variation of a stable heuristic like win-stay, lose-shift, suggesting that it is not one heuristic, but a result of two independent heuristic processes that separately react to reward and loss. Consistent with this, participants respond more quickly to losses than wins (Dyson et al., 2018) and exhibit fairly distinct EEG signatures when responding to different game outcomes (Dyson et al., 2020; Dyson et al., 2018). Further, it appears that win-stay behavior may not arise as consistently as lose-shift (Dyson et al., 2020; Dyson et al., 2016) and may be more vulnerable to fluctuations in game rewards (Forder & Dyson, 2016). Whether win-stay and lose-shift reflect a single mechanism or not, this class of accounts suggests that human behavior in the rock, paper, scissors game is best explained by a conjunction of stable heuristics.

While win-stay, lose-shift and other heuristic strategies may offer people a simple decision process, they are also insufficient to explain human behavior in repeated games of rock, paper, scissors. For one, dependencies in people's move choices extend beyond such heuristics to a variety of other complex sequential regularities which cannot be as easily accounted for (Brockbank & Vul, 2020). Second, an emphasis on heuristics as a basis of people's decision making in repeated RPS interactions fails to address the ways in which people exhibit more dynamic, adaptive behavior, such as exploiting biases in their opponent's choices (Kangas et al., 2009; Lie et al., 2013; West & Lebiere, 2001).

Recent results suggest people are trying to outwit their opponents

A complete account of human behavior in repeated MSE games like rock, paper, scissors should accommodate the adaptive character of people's decision making over many interactions. Consider, for example, playing repeated games with an opponent that simply plays "rock" over and over. Here, subjective randomness or win-stay, lose-shift responding would be surprising. Though trivial, this illustrates a critical underlying dynamic in repeated MSE games: *Optimal play depends on the predictability of the opponent.* Heuristics or subjectively random behavior may be adaptive against an unexploitable opponent, and may serve as a useful fallback when one is losing, but they are not the best policy when facing a fallible opponent. In large scale algorithmic RPS tournaments, random strategies often under-perform precisely because they fail to detect stable dependencies in their opponent's moves that could be exploited (Billings, 2000a, 2000b).[1]

Despite its intuitive appeal, the role of *adaptive, adversarial* reasoning in repeated RPS interactions has been largely overlooked in the prior literature. Most empirical studies of rock, paper, scissors behavior pair participants either against automated opponents employing a random strategy (Dyson et al., 2020; Dyson et al., 2018; Dyson et al., 2016; Forder & Dyson, 2016; Gallagher et al., 2002; Kangas et al., 2009; Lie et al., 2013; Stöttinger, Filipowicz, Danckert, et al., 2014), or against a shuffled group of human opponents (Frey & Goldstone, 2013; Hoffman et al., 2012; Wang & Xu, 2014; Wang et al., 2014; Xu et al., 2013). In both cases, participants cannot adapt to the dependencies of their opponent. Random computer choices are simply unexploitable, while random assignment of opponents ensures that sequential choices are independent and identically distributed, and thus equally unexploitable through more sophisticated adversarial reasoning. Thus,

[1] Successful algorithmic strategies in a recent Kaggle RPS tournament highlight this dynamic: `https://www.kaggle.com/c/rock-paper-scissors`.

these results cannot address whether decision-making over repeated interactions, including the sequential regularities observed in prior empirical work, may result from an effort to *outwit* one's opponent.

What happens when people play against opponents that *are* exploitable, such as stable human adversaries? A handful of recent studies asking this question yield behavior consistent with flexible, adaptive reasoning, rather than simple heuristics or subjective randomness. First, in repeated interactions with opponents that exhibit a strong bias towards certain moves, people often show an above-chance capacity to exploit the opponent (Kangas et al., 2009; Lie et al., 2013) consistent with basic reinforcement learning mechanisms (Sepahvand et al., 2014). Notably, this adaptability appears to be limited to very strong opponent biases, even over many trials (Danckert et al., 2012; Filipowicz et al., 2016; Stöttinger, Filipowicz, Marandi, et al., 2014). However, efforts to outwit a stable opponent extend beyond reinforcement learning and draw on more structured pattern recognition abilities when opponent behavior is more nuanced. Stöttinger, Filipowicz, Danckert, et al. (2014) find that people adapt to bots that exhibit a Cournot Best Response transition strategy, but their ability to do so is limited by prior exposure to an opponent with a simple move bias, suggesting a strong role of context in adversarial reasoning. West and Lebiere (2001) provide a relatively thorough investigation of people's ability to adapt to neural network opponents with a memory for various numbers of previous moves, showing that people are reliably able to beat a *lag1* opponent whose moves are primarily based on the previous move, but behave more similarly to a *lag2* opponent that draws on the two previous rounds. However, recent work has found that people can detect even more complex transition and outcome-dependent transition strategies over many rounds (Brockbank & Vul, 2023; Dyson et al., 2020; Dyson et al., 2018). Finally, results in Brockbank and Vul (2021) indicate that when paired with opponents that exploit regularities in participants' *own* move choices, people are able to *counteract* such exploitation for simpler behavioral dependencies. Taken together, these results suggest that over many RPS interactions with a stable opponent, people are highly attuned to the structured dependencies which make players themselves and their opponents exploitable.

In sum, recent results suggest that people's behavior over many rounds against a potentially exploitable opponent can be explained by the desire to outwit that opponent, rather than merely attempting to respond randomly or relying on stable heuristics. But how flexible is this ability, and what are its limitations? What sorts of hypotheses about behavioral structure can people entertain and track on the basis of an opponent's sequential decisions? Addressing these questions requires characterizing the space of uniquely identifiable strategies that may be exploited, and estimating whether people attend to these regularities when playing repeated rounds of RPS. The rest of the paper focuses on these technical challenges.

1.3 RPS behavior reveals structure of adversarial reasoning

Human behavior during repeated interactions in mixed strategy games like rock, paper, scissors may be explained by ongoing attempts to *outwit one's opponent*. However, it remains an open question *how* people are able to adapt to regularities in an opponent's behavior. What kind of dependency structures can people detect and respond to? Prior work has examined the ways that different sequential patterns in RPS can be categorized (Dyson, 2019). Building on these results, we begin by providing an overview of how the complex dependencies observed in people's move decisions are structured and show how people's exploitability along these dimensions can be quantified. We then demonstrate how such measures can be used to explore which behavioral regularities people successfully exploit against a stable opponent. We apply these methods to experimental data from prior work by Brockbank and Vul (2020) to explore how well different sequential regularities predict people's move decisions and the degree to which they successfully exploit regularities in an opponent's behavior. In this way, we show that behavior in the rock, paper, scissors game offers novel insights into how people perform adaptive, adversarial reasoning.

1.3.1 Sequential dependencies in RPS can be systematically described

Individual dependencies

The level at which people are able to outwit their opponents (i.e., the scope of their adversarial reasoning abilities) is reflected in the structure and complexity of the sequential dependencies they can detect and exploit, and how much they do so over many rounds. How can we define this structure, and how do we then assess whether these dependencies are exploited by a savvy player? In the rock, paper, scissors game, the space of exploitable dependencies can be described in increasing order of complexity based on the number of prior events that impact a player's move choices (Brockbank & Vul, 2020; Dyson, 2019). In other words, sequential dependencies in a player's RPS moves are expressible in terms of how the probability of a particular *decision*—either a move selection or a transition between moves—is statistically impacted by some form of previous *event*: the player's own previous move, their opponent's previous move, etc. If a player or bot is behaving randomly, the probability of any decision will be equal no matter what previous event is considered; every move or transition is just as likely given every previous move or outcome. However, to the degree that a player's behavior is exploitable, they will exhibit non-uniform move or transition probabilities conditioned on a particular event, such as their previous move. The greater the departure from a uniform distribution conditioned on the prior event, the more exploitable a player is, i.e., the more they exhibit this dependency. Broadly, the more prior events required to evaluate the dependency, the more complex it is. Questions about a person's adversarial reasoning abilities in RPS therefore come down to measuring whether and how much they can recognize these dependencies in their opponent.

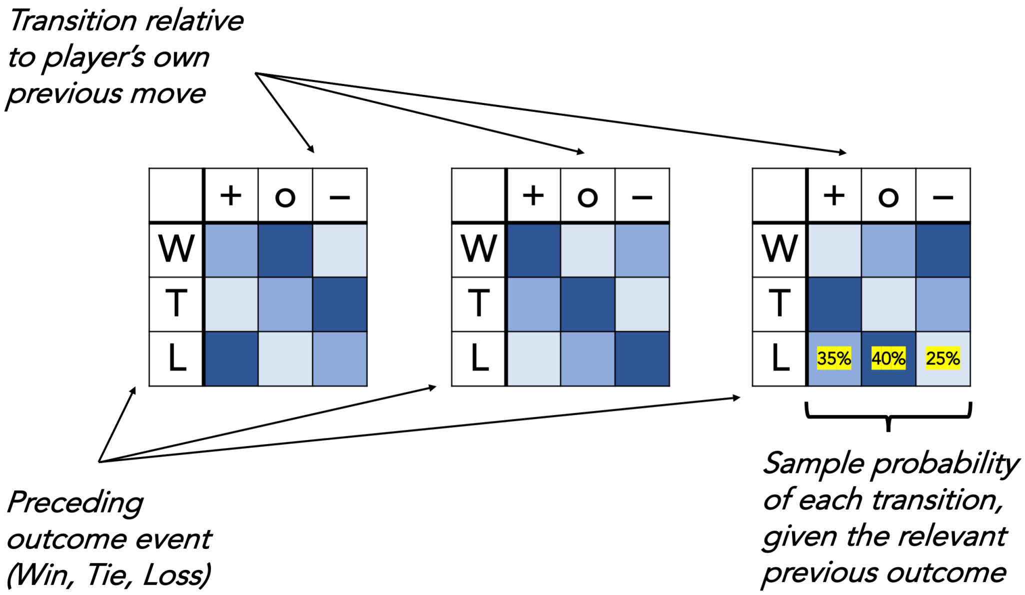

Transition relative to player's own previous move
+
o
–
W
T
L
35%
40%
25%
Preceding outcome event (Win, Tie, Loss)
Sample probability of each transition, given the relevant previous outcome

Figure 1.2. Schematic of dependency variants exhibited during rock, paper, scissors play. Above are three distinct versions of an outcome-dependent transition dependency like win-stay, lose-shift. Shaded squares indicate gradations in the probability of a given transition (column) given each prior event (row).

To illustrate, the tables in Figure 1.2 show how *outcome-based* transition dependencies like win-stay, lose-shift can be represented. Here, each state of a dependency *event* like previous outcome is given a unique row on the left side of the table. The dependencies in Figure 1.2 have a row for each possible outcome from the previous round—win (W), tie (T), and loss (L)—but a simpler dependency based on, e.g., one's own previous move might instead have a row for "rock", "paper", and "scissors". Each column indicates a possible *decision* based on that row-wise dependency event. In Figure 1.2, these decisions are move transitions: shift up (+), stay (0), or shift down (−). Once again, a simple dependency in which move choices are based on one's own previous move could be expressed with possible move decisions (“rock”, “paper”, “scissors”) in each column instead of transitions. Each cell in the tables in Figure 1.2 then represents the probability that the player chooses the action in the cell’s column following the dependency event in the corresponding row. If players did not exhibit any dependency on a row-wise outcome, the probabilities in each cell in that row would be 1/3, signaling that each transition (column value) is equally likely given that row value. However, the more a player exhibits a particular dependency, the greater the disparity between their transition probabilities given each possible outcome. This encoding of patterned behavior therefore allows us to express each unique *class* of dependencies that a player could exploit in their opponent through the choice of different row-wise events and column-wise actions. The ability to express RPS dependencies in this way is not limited to outcomes affecting transition choices, as in Figure 1.2, but applies at every level of behavioral complexity. This structure for expressing classes of sequential patterns therefore provides a formal mechanism for outlining the *hypothesis space* of behavioral regularities people exhibit and can adapt to. In the next section, we discuss this space, in particular, the relationship between different dependencies.

Combining dependencies

Critically, the various classes of sequential dependencies that a player can exhibit in their move choices are not independent, but rather are arranged in an expressive hierarchy. Dependencies exhibited at one level will affect other levels that rely on the same information. For example, a player’s distribution over moves given the previous move subsumes their marginal distributions over transitions and moves—any pattern in their overall move or transition distributions will be reflected in the distribution of moves given the previous move. Why is this important when describing people’s adaptive behavior? If a player exhibits a tendency toward a particular move following each previous move, this will in part reflect any lower level biases in their moves and transitions. Describing their behavior as following a strategy of gravitating toward particular moves after each previous move must factor in the degree to which they are simply favoring some moves or transitions. Similarly, if a player is able to exploit an *opponent* seemingly on the basis of regularities in the opponent's moves following each previous move, we want to know that they are not primarily sensitive to simpler dependencies in the opponent's transition or move base rates. Broadly, the dependency signal for a given dependency structure will include the dependency signal from its lower level subsidiaries.

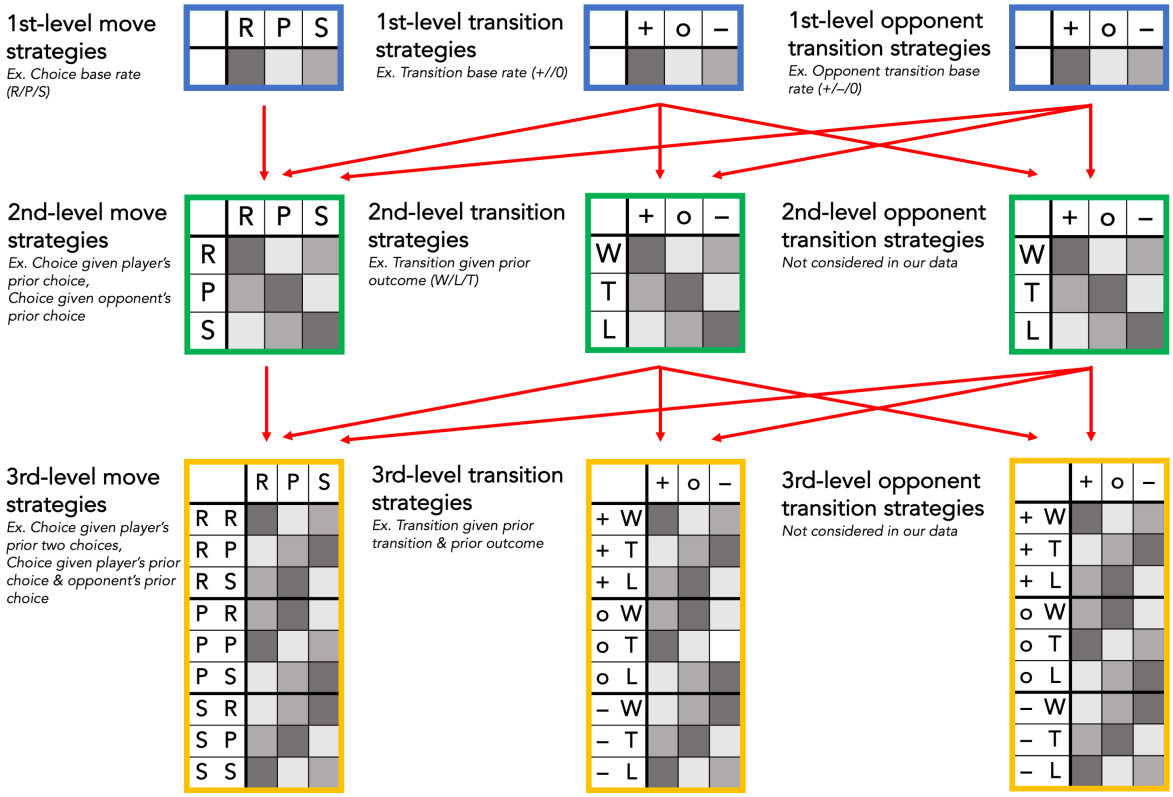

1st-level move strategies
Ex. Choice base rate (R/P/S)
R P S
1st-level transition strategies
Ex. Transition base rate (+//0)
+ o –
1st-level opponent transition strategies
Ex. Opponent transition base rate (+/–/0)
+ o –
2nd-level move strategies
Ex. Choice given player's prior choice, Choice given opponent's prior choice
R P S
R
P
S
2nd-level transition strategies
Ex. Transition given prior outcome (W/L/T)
+ o –
W
T
L
2nd-level opponent transition strategies
Not considered in our data
+ o –
W
T
L
3rd-level move strategies
Ex. Choice given player's prior two choices, Choice given player's prior choice & opponent's prior choice
R P S
R R
R P
R S
P R
P P
P S
S R
S P
S S
3rd-level transition strategies
Ex. Transition given prior transition & prior outcome
+ o –
+ W
+ T
+ L
o W
o T
o L
– W
– T
– L
3rd-level opponent transition strategies
Not considered in our data
+ o –
+ W
+ T
+ L
o W
o T
o L
– W
– T
– L

Figure 1.3. Schematic for quantifying complexity of dependencies exhibited during rock, paper, scissors play. On the left are three levels of increasing complexity for regularities in players' move choices. In the middle and right columns are equivalent complexity levels for dependencies players exhibit in their transitions between moves, either relative to their own previous move, or relative to the opponent's. The arrows illustrate the hierarchical relationship across these regularities, indicating for example how second-level move dependencies carry some of the dependency signal captured by first-level move and transition dependencies.

The schematic in Figure 1.3 shows the inheritance relationship among increasingly complex sequential move and transition dependencies. As the dependencies become more complex, they inherit from a greater number of simpler regularities. While this does not show the full space of *possible* regularities (such a space is technically infinite), we include any behavioral dependencies that have been observed in prior work (i.e., all of those discussed in our review of existing literature) or in previous attempts to frame these structures (Brockbank & Vul, 2021; Dyson, 2019). For researchers attempting to quantify how much people are exploitable or are successfully exploiting opponents on the basis of these dependencies, this structure poses a credit assignment problem: how to identify when a dependency is being exploited above and beyond the lower level dependencies it is based on? The key to attributing behavior at the right level of complexity is to use this hierarchical dependency structure when evaluating the regularities in people's move choices. In other words, to untangle the unique contribution of a higher-order dependency structure from the exploitability arising from its subsidiaries, we partial out the subsidiary dependencies based on the relationships in Figure 1.3. This allows us to ask how much each dependency contributes to explaining individual behavior. As we show below, this logic can be applied not only to estimating a given player's level of exploitability within a given structure, but also to estimating how much this dependency is exploited by their opponent.

1.3.2 Quantifying how much people exhibit and exploit sequential dependencies

In the previous section, we showed that the exploitable dependencies people exhibit over repeated rounds of rock, paper, scissors can be described in terms of how events like previous moves or outcomes impact the probability of subsequent move decisions. We further showed that the relationship among different dependencies of this sort prevents us from treating them independently without correcting for the shared structure across dependencies. How then can we quantify *how much a player exhibits a given dependency* and, relatedly, *how much their opponent is able to exploit it*?

Measuring exploitability with information gain

We measure how predictable a player's behavior is subject to a particular dependency via conditional entropy and information gain. In rock, paper, scissors, the player has three choices, $a_{1-3} \in A$. This action space A can either represent the move choices ("rock", "paper", and "scissors"), or the transitions (+, −, 0) relative to the player's previous move or relative to the opponent's previous move (the set of transitions encodes additional information about either the player or their opponent's previous move but is otherwise the same). A player's propensity to make some choices more than others in a given context c (i.e., how exploitable they are in this context), can be summarized as the probability distribution $P(a_i \mid c)$. The Shannon entropy (Shannon, 1948) of the distribution over those choices describes how unpredictable they are:

$$H(A \mid c) = -\sum_{i=1}^{3} P(a_i \mid c) \log_2 P(a_i \mid c)$$

,

and will take on a value, in bits, between 0—for completely deterministic behavior, where one of the three actions is always chosen in a given context—and $\log_2 3$ for uniform behavior where all three actions are equally likely.

In the base case, where the context, c is an empty set, this definition is sufficient, and reduces to entropy over actions $H(A)$. However, for all non-trivial contexts, we calculate the Shannon entropy for each possible state in the context and average over them. For instance, a strategy such as "win-stay, lose-shift" describes a distribution over self-transitions that varies with context defined as the outcome of the preceding round. Our entropy calculation must factor in the full *partition* over contexts $\mathbb{C}$ that a dependency structure imposes. In the case of win-stay, lose-shift, the relevant dependency structure defined by the context partition is: $\mathbb{C} = \{\text{win}, \text{loss}, \text{tie}\}$. The unpredictability of choices given a context partition is therefore given by the conditional entropy marginalized over the contexts in that dependency structure:

$$H(A \mid \mathbb{C}) = \sum_{c \in \mathbb{C}} P(c) H(A \mid c)$$

.

To characterize how much behavioral regularity may be captured via a particular dependency structure defined by the partition over contexts ($\mathbb{C}$), we ask how much information is gained about actions by taking that dependency structure into account. Specifically, we can subtract the conditional entropy under that dependency structure from a uniform distribution over choices, to calculate the information gained by using that dependency structure to predict a player's moves or transitions:

$$I(A \mid \mathbb{C}) = \log_2 3 - H(A \mid \mathbb{C})$$

.

Intuitively, this measure quantifies the improvement gained by predicting a player's moves or transitions using a particular dependency relative to a random baseline. Large information gain for a given dependency structure suggests that a player is highly exploitable via that dependency. Low values suggest that their behavior is not easily distinguished from random choices given the prior events in $\mathbb{C}$.

While information gain provides an intuitive measure for how much a player exhibits a particular dependency, it fails to reflect the hierarchical structure of dependencies described previously. In other words, the information gain associated with a given dependency structure will not capture just the information *unique* to that structure. For instance, if a player shows a bias toward choosing "rock", that predictable dependency will also show up in the information gain over each move conditioned on the previous move. To uniquely identify the information gained for a particular dependency structure, we must consider the hierarchical structure of different dependencies shown in Figure 1.3.

Given the hierarchical relationship among dependency structures in Figure 1.3, we can define an operation $\Phi(\mathbb{C})$ which yields all the upstream nodes (parents, grandparents, etc.) of a given dependency structure. For instance, the dependency structure capturing the tendency to choose “rock”, “paper”, or “scissors” given one’s previous choice has two parents: an overall move bias to choose “rock”/“paper”/“scissors”, and a preference for particular self-transitions (+/−/0). Using this, we can calculate a *corrected* information gain for a particular dependency structure by subtracting the information gained from the parent dependency structures:

$$I^*(A \mid \mathbb{C}) = I(A \mid \mathbb{C}) - \sum_{\mathbb{B} \in \Phi(\mathbb{C})} I^*(A \mid \mathbb{B})$$

.

This calculation yields a measure of the information about actions that can be *uniquely* captured in a given dependency structure. The ability to attribute sequential patterns in behavior to a particular dependency structure is critical for understanding the cognitive processes underlying adversarial reasoning in the rock, paper, scissors game. Prior work has shown that certain patterns of outcome-based transition behavior (i.e., win-stay, lose-shift) are *isomorphic* to much simpler patterns of Cournot best responding when a player’s self-transitions are re-cast as transitions relative to their opponent’s previous move (Dyson, 2019). Because of this isomorphism, conclusions about whether a savvy player is exploiting complex outcome-based patterns in their opponent, or is simply sensitive to the pattern of Cournot transition responses may be ambiguous. Here, by correcting the information gain for a given dependency structure to reflect all upstream parents, we can identify the extent to which people exhibit dependencies of a certain complexity, without being misled by the possibility of a complex dependency being mimicked by a simpler one. More broadly, this provides a means of quantifying how much players exhibit rich and complex patterns in their move choices over many rounds. Answering this allows us to then address questions at the heart of adversarial reasoning in the rock, paper, scissors game: which behavioral patterns do people exploit in their opponents? Generally, what is the relationship between how much people exhibit a particular behavioral regularity and how much their opponents are able to exploit it?

Measuring how much players are exploited with expected win count differentials

To understand the relationship between a player's exploitable behavior patterns and whether their opponent in fact uses these patterns to their advantage, we extend the information gain measure described previously to reflect the outcomes that might be *expected* by fully exploiting a given dependency in a player's moves. Intuitively, the level at which a player's decisions over repeated rounds are exploitable can be thought of as the number of games their opponent could expect to win by taking advantage of the patterns their choices exhibit. We refer to this as the *expected win count differential* for a given dependency structure. The win count differential is simply the number of games that one player wins over the course of many rounds minus the number of games won by their opponent. A positive win count differential for one player indicates that they were able to win more often than their opponent and higher win count differentials indicate more successful exploitation of the opponent. The *expected* win count differential, then, captures how much advantage a player could theoretically obtain by choosing moves which maximally exploit a particular dependency in their opponent's moves. Given a non-uniform (exploitable) distribution over an opponent's actions $P(a_i \mid c)$, a player's expected win count differential for a given action a_j is equal to $\sum_i P(a_i \mid c) \cdot v(a_i, a_j)$, where $v(a_i, a_j) \in \{-1, 0, 1\}$ is the outcome of playing a particular move a_j against the opponent's move a_i: increasing the player's win count by 1, decreasing by 1, or tying for a change of 0. Given this, the player has an *optimal* action j^* that maximizes their expected win count differential over all possible opponent moves:

$j^* = \arg\max_j \sum_i P(a_i \mid c) \cdot v(a_j, a_i)$.

This optimal choice in turn yields an expected win-count differential of: $\mathbb{E}[v \mid c] = \sum_i P(a_i \mid c) \cdot v(j^*, a_i)$. And averaging over all contexts (for example, the set of all previous moves by the player), this yields:

$$\mathbb{E}[v \mid \mathbb{C}] = \sum_{c \in \mathbb{C}} \mathbb{E}[v \mid c] P(c)$$

.

The expected win count differential for a given dependency context $\mathbb{C}$ captures how exploitable a player is along that dimension, much like the information gain measure described previously. In fact, the difference between the expected win count differential and the information gain for a particular dependency structure is often small, since lots of information in a given dependency will translate directly into expected win count differentials. However, not all low-entropy distributions are equally exploitable. For instance, a player that chooses their moves with the distribution 60% "rock", 30% "paper", and 10% "scissors", can be exploited to achieve an average win count differential (per round) of 0.5 by playing "paper". Meanwhile, a move distribution of 60% "rock", 10% "paper", and 30% "scissors" only yields an expected win count differential of 0.3 (by playing "scissors"; playing "paper" yields an expected win count differential of only 0.2). These two distributions have the same entropy and information gain, but one is nearly twice as exploitable as the other, in terms of the achievable win count differential. Thus, expected win count differential tells us not just how much information is available at a given dependency structure, but how exploitable such information is.

As a measure of how exploitable a player's behavior is, expected win count differential also enables us to investigate the relationship between how much a player's opponent could theoretically exploit patterns in their behavior, and how much their opponent actually did so. This is because expected win count differentials can be directly compared to *observed* win count differentials in dyads, indicating whether regularity at a particular dependency structure might explain the observed pattern of advantage seen in a pair of players. Given a set of many repeated RPS games between pairs of stable opponents, we can use each player's level of exploitability for a given dependency—their *expected* win count differential—as predictors in a regression over the true win count differentials in each dyad. This provides a first approximation of how much of the variance in empirical win count differentials can be explained by the different ways that players exhibit exploitable behavior across many dyads.

However, this approach faces the same fundamental challenge as the uncorrected information gain measure described earlier; expected win count differentials for different behavioral regularities will be influenced by the rich interdepencence of these regularities shown in Figure 1.3. Thus, predicting empirical win count differentials using raw expected win count differentials fails to accommodate the role of lower level dependencies in higher level expected win count differentials. In this context, to correct expected win count differentials for upstream dependencies, we cannot simply subtract them, as we can for information gain. Instead, we correct for the hierarchy in Figure 1.3 within the observed win count differential regression itself. To illustrate, when predicting observed win count differentials across experimental dyads, we only use the simplest dependencies in Figure 1.3 as direct predictors. To partial out the role of these lower level dependencies in more complex dependencies, we include the *residuals* from separate regressions of expected win count differentials for each higher level dependency predicted by expected win count differentials for the dependencies they inherit from. For example, a player's level of exploitability using *2nd-level move strategies* in Figure 1.3, such as their choice given their prior choice, can be predicted based on their exploitability using *1st-level move strategies* (base rate of “rock”, “paper”, and “scissors”) and *1st-level transition strategies* (base rate of $+$, $-$, and 0 transitions). The residuals from this prediction using expected win count differentials indicate how much of the variance in a given 2nd-level move strategy *cannot* be accounted for by the 1st-level strategies. These residuals can then serve as predictors for the 2nd-level variables in the original regression of observed dyad win count differentials. In this manner, we can isolate the unique dependency arising at a certain level of behavior, rather than attributing lower level dependencies to the more abstract, higher-order structure.

To summarize, we have argued that behavior in repeated games of rock, paper, scissors provides a window into how people perform the sort of adaptive, adversarial reasoning that allows them to outwit a stable opponent. We first showed that a player’s exploitable behavior—patterns that their opponent might use to their advantage—contains structure illustrated in their conditional move or transition probabilities subject to various contingencies like their previous move. We further showed how these regularities are hierarchically arranged. Given this, we next showed how a player’s exploitability, i.e., the degree to which they exhibit a given dependency structure, can be quantified using measures of information gain and expected win count differential. The former indicates exactly how much *signal* is contained in a player’s patterned behavior, and the latter incorporates the way this signal can be exploited. Finally, we showed how the level of exploitability that a player exhibits can be used to investigate which sources of exploitability contribute to the observed pattern of players exploiting their opponents, thus providing clues about the underlying nature of people’s adversarial reasoning in this setting. In the next section, we show how these measures can be applied to empirical data to explore the flexibility and limitations of people’s ability to outwit an opponent.

1.4 Adversarial reasoning in RPS relies on detecting simple regularities

In the previous section, we showed how sequential regularities in people's move decisions in rock, paper, scissors can be formally described and quantified. This might serve as the basis for a more precise characterization of the dependencies people exhibit in their own behavior in adversarial settings, as well as the patterns they can detect and exploit in opponents. In other words, this framework offers a unified view of the decision making biases shown in rock, paper, scissors move choices (Baek et al., 2013; Brockbank & Vul, 2020; Dyson et al., 2016; Wang et al., 2014), and the complexities of modeling *opponent* behavior in the same setting (Brockbank & Vul, 2023; Brockbank & Vul, 2021; Dyson et al., 2018; Stöttinger, Filipowicz, Danckert, et al., 2014; West & Lebiere, 2001).

Here, we show how the measures from the previous section can be applied to empirical data from a set of rock, paper, scissors dyads. Brockbank and Vul (2020) paired 116 participants into stable dyads and collected data for 300 rounds of rock, paper, scissors in each dyad. Because participants in this experiment were playing with the same opponent for 300 consecutive rounds, players had ample time to try and learn sequential patterns in their opponent's moves. Indeed, the authors find that the distribution of empirical win count differentials across the 58 dyads is overall significantly larger than would be expected under random play, suggesting that players found ways to outwit their opponents. How did some participants perform the adaptive, adversarial reasoning necessary to gain a steady advantage over their opponents? Here, we attempt to answer this question using the measures outlined in the previous section. We first examine the average information gain for a range of sequential dependencies proposed in Brockbank and Vul (2020) to quantify how much participants exhibited exploitable patterns. Next, we explore the relationship between observed win count differentials and expected win count differentials to assess which patterns best explain participants' ability to outwit their opponents.

1.4.1 People exhibit complex behavioral dependencies

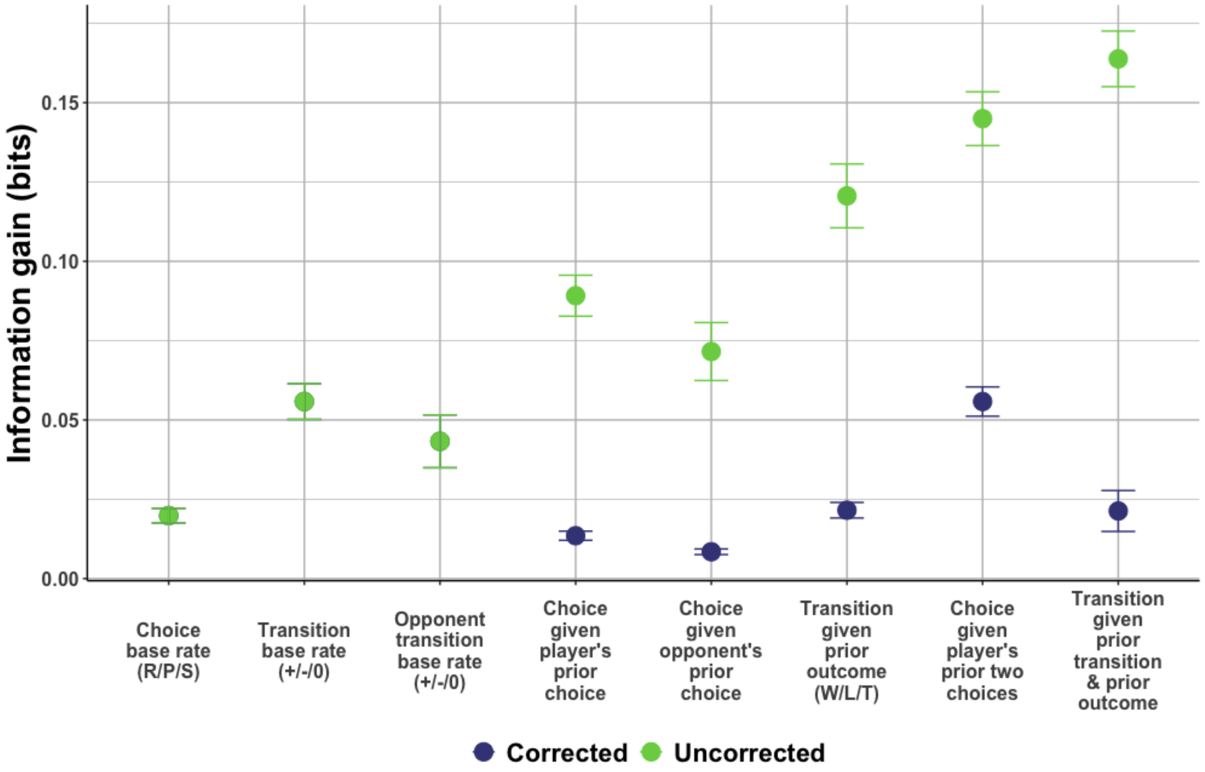


Figure 1.4. Change in average information gain (bits) as a result of incorporating the hierarchical structure in Figure 1.3. The information gain reflects how exploitable individuals were for each of the dependencies shown. For more complex dependencies, individual exploitability decreases when corrected for simpler low-level dependencies. Error bars show one SEM.

The data from Brockbank and Vul (2020) suggest that across 300 rounds, people exhibit stable predictable behaviors that might form the basis of exploitation by their opponents. Here we ask how predictable their behavior was for a range of sequential regularities. In particular, we ask how the Shannon entropy over RPS choices for a given player is reduced when conditioning on some prior dependency. As outlined above, the reduction in entropy compared to chance behavior represents the information gain from taking each dependency structure into account. Figure 1.4 shows average information gain across participants for eight different dependency structures that increase in complexity from left to right. We plot the "Uncorrected" information gain values for each dependency alongside the “Corrected” information gain to account for the hierarchical structure of these dependencies as described previously. Larger information gain (in bits) indicates a greater level of predictability for that particular dependency. The uncorrected values show a steady increase in information gain as the complexity of the dependency increases on x, suggesting greater and greater predictability for more complex sequential patterns. However, the corrected values suggest that some of this increase can be attributed to higher level patterns carrying signal from lower level ones. Nonetheless, the complex dependencies at the right retain some signal even after correction, providing evidence that people’s move choices are exploitable using a range of sequential patterns that vary in their complexity.

1.4.2 Players exploit simple behavioral dependencies in their opponents

Across repeated games of rock, paper, scissors with a stable opponent, Brockbank and Vul (2020) show that some players are able to reliably outwit their opponents. But among dyads that exhibit higher win count differentials, what kinds of regularities in one player’s move choices form the basis of this exploitation by their opponent? In other words, which dependencies do people successfully exploit?

As described in the previous section, we can begin to address this question by exploring the relationship between the observed win count differentials in each dyad and the average *expected* win count differentials in each dyad for each of the sequential dependencies that players may have relied on to exploit their opponent. Critically, we correct for the hierarchical relationship among dependencies using the residuals from separate regressions for complex dependencies where some of the predictability may derive from simpler underlying dependencies. Using the dyad results from Brockbank and Vul (2020) as the basis for this regression, we find that expected win count differential based on transition dependencies (the *transition base rate (+/-/0)*) and opponent previous move dependencies (*player's choice given opponent's prior choice*) are both significant predictors of empirical win count differentials in each dyad (Transition: $\hat{\beta} = 0.19$, $p = 0.027$; Opponent previous choice: $\hat{\beta} = 0.45$, $p = 0.015$). In other words, the degree to which players exploit their opponents over 300 rounds is best explained by simple biases that players in the dyad exhibit toward particular transitions, as well as regularities in player moves given their opponent's previous move.

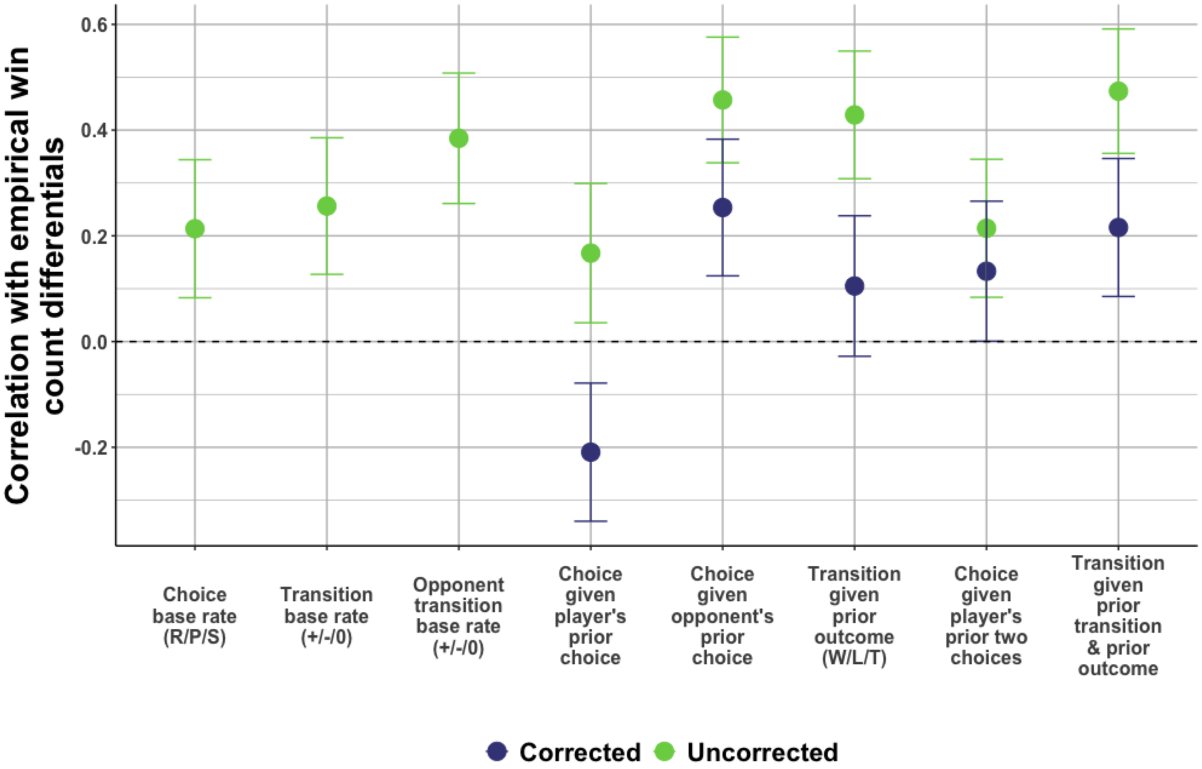


Figure 1.5. Change in the relationship between expected win count differential for each behavioral dependency and empirical win count differentials as a result of incorporating the hierarchical structure in Figure 1.3. For more complex dependencies, the role they play in exploitation among dyads decreases when we factor in the role of lower level dependencies. Error bars show one SEM.

But what might the regression look like if we *did not* correct for the hierarchical structure of the dependencies? Figure 1.5 plots the correlation between *expected* win count differentials—how much players in each dyad exhibited each dependency—and true win count differentials, i.e., how much players in each dyad exploited their opponents overall.

Critically, we first plot these correlations using the expected win count differentials for each dependency ("Uncorrected" correlations), and then substitute them for the the residuals as described in the previous section (the "Corrected" correlations). Figure 1.5 illustrates the importance of this correction; revised correlations are broadly lower across the board, but especially for the most complex dependencies on the right. Therefore, incorporating the hierarchical structure of the dependencies into the correlation shows that people's use of complex regularities when exploiting their opponent may in fact draw heavily on simpler, low-level behavioral patterns.

1.5 Discussion

Here we argued that games like rock, paper, scissors offer a precise and tractable way to study adaptive, adversarial reasoning. We started with the observation that human play in simple cyclic-dominance games, such as matching pennies or rock, paper, scissors, systematically deviates from the mixed strategy Nash Equilibrium of purely random play. In particular, people exhibit a range of sequential regularities in their move choices that are most consistent with an intuitive, but understudied account: people are constantly trying to *outwit* their opponents, and behavioral dependencies arise from such adaptive reasoning.

How can we make sense of the behavioral regularities that emerge as a result of adaptive reasoning in the rock, paper, scissors game? Building on prior work exploring the cognitive and computational resources required to identify such dependencies (Dyson, 2019), we outline a schema for formally describing the ways that rock, paper, scissors behavior can reflect stable patterned regularities. We show that the predictability and subsequent exploitability of a given dependency can be precisely quantified using measures of conditional entropy and *expected* win count differentials. Prior work in this space raised important concerns about the identifiability of complex dependency structures in a player's behavior due to isomorphisms between different patterns in behavior which make distinctly different cognitive demands of an adaptive opponent (Dyson, 2019). To overcome this challenge, we introduce analytical techniques that can correct for the hierarchical inheritance structure among different dependencies, and can thus identify both the extent to which people exhibit, and exploit, complex behavioral patterns.

Finally, we validate our approach by applying the proposed measures of exploitability and adversarial reasoning to a large empirical dataset comprised of repeated rock, paper, scissors games between a set of stable dyads from Brockbank and Vul (2020). Our results show that incorporating the hierarchical structure of sequential dependencies into analysis of human behavior allows for a clear description of how each dependency is reflected in individual decisions. Concretely, our results offer two key findings which highlight the value of repeated rock, paper, scissors interactions in understanding human adaptive reasoning capacities. First, we show that over many rounds against a stable opponent, people exhibit a range of exploitable dependencies, including some that reflect a high level of complexity. These however, are attenuated by the expression of simpler dependencies. Next, we show that despite the range of predictable behavior patterns in people's decisions, their opponents largely fail to exploit these same dependencies. Instead, people rely on simple transition and previous move dependencies in order to outwit their opponents, an intuitive finding that our results provide concrete, quantitative support for.

The current results show that the rock, paper, scissors game can be fruitfully used to study the flexibility of human adversarial reasoning. In particular, we show how people's behavior across repeated interactions reveals the limits of our capacity to detect and adapt to sequential behavior patterns. Critically, rock, paper, scissors presents just one avenue by which these and other similar questions can be addressed. Applying a similar approach to other mixed strategy equilibrium games, or even a broader set of strategic interactions altogether, may reveal further insights about adversarial reasoning. In particular, one interpretation of the current results is that the failure to exploit more complex dependencies arises from limits in memory. Prior work has considered the impact of memory length on strategic behavior in a range of domains including RPS (Posch, 1999; West & Lebiere, 2001); the current results may open the door to a more precise account of such resource limits in adversarial reasoning.

Together, our results show how the simple game of rock, paper, scissors can support a quantitative perspective on the rich adaptive reasoning and opponent modeling that underlies human competition. What kinds of complex, patterned behavior can people detect and adapt to in strategic settings, and how does dyadic behavior reflect exploitation of these patterns across repeated interactions? We hope that our framework for constructing and analyzing dependencies in rock, paper, scissors allows researchers to better characterize human adaptive adversarial capacities.

1.6 Acknowledgments

I thank my co-author Edward Vul for his unparalleled sagacity and peerless equanimity during the daunting preparation of this manuscript. I also thank the additional members of my qualifying exam committee for their feedback on an earlier version of this manuscript: Judith Fan and Craig McKenzie.

Chapter 1, in full, is a reprint of material as it appears in Brockbank, E., & Vul, E. (2021). Formalizing opponent modeling with the rock, paper, scissors game. *Games*, *12*(3), 70. An earlier version of the project was published as Brockbank, E., & Vul, E. (2020). Recursive adversarial reasoning in the rock, paper, scissors game. In S. Denison., M. Mack, Y. Xu, & B.C. Armstrong (Eds.), *Proceedings of the 42nd Annual Conference of the Cognitive Science Society* (pp. 1015-1021). Cognitive Science Society. The dissertation author was the primary investigator and author of this material.

References

Aczel, B., Bago, B., & Foldes, A. (2012). Is there evidence for automatic imitation in a strategic context? *Proceedings of the Royal Society B: Biological Sciences*, *279*(1741), 3231–3233.

Allesina, S., & Levine, J. M. (2011). A competitive network theory of species diversity. *Proceedings of the National Academy of Sciences*, *108*(14), 5638–5642.

Axelrod, R. (1984). *The evolution of cooperation*. Basic Books.

Baek, K., Kim, Y. T., Kim, M., Choi, Y., Lee, M., Lee, K., Hahn, S., & Jeong, J. (2013). Response randomization of one-and two-person rock-paper-scissors games in individuals with schizophrenia. *Psychiatry research*, *207*(3), 158–163.

Bar-Hillel, M., & Wagenaar, W. A. (1991). The perception of randomness. *Advances in applied mathematics*, *12*(4), 428–454.

Billings, D. (2000a). The first international roshambo programming competition. *ICGA Journal*, *23*(1), 42–50.

Billings, D. (2000b). Thoughts on roshambo. *ICGA Journal*, *23*(1), 3–8.

Brockbank, E., & Vul, E. (2023). Rock, paper, scissors play reveals limits in adaptive sequential behavior. *Manuscript submitted for publication.*

Brockbank, E., & Vul, E. (2020). Recursive adversarial reasoning in the rock, paper, scissors game. In S. Denison, M. L. Mack, Y. Xu, & B. C. Armstrong (Eds.), *Proceedings of the 42nd annual conference of the cognitive science society* (pp. 1015–1021). Cognitive Science Society.

Brockbank, E., & Vul, E. (2021). Humans fail to outwit adaptive rock, paper, scissors opponents. In T. Fitch, C. Lamm, H. Leder, & K. Teßmar-Raible (Eds.), *Proceedings of the 43rd annual conference of the cognitive science society* (pp. 1740–1746). Cognitive Science Society.

Brown, J. N., & Rosenthal, R. W. (1990). Testing the minimax hypothesis: A re-examination of o'neill's game experiment. *Econometrica: Journal of the Econometric Society*, *58*(5), 1065–1081.

Budescu, D. V., & Rapoport, A. (1994). Subjective randomization in one- and two-person games. *Journal of Behavioral Decision Making*, *7*(4), 261–278.

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

Cason, T. N., Friedman, D., & Hopkins, E. (2014). Cycles and instability in a rock-paper-scissors population game: A continuous time experiment. *Review of Economic Studies*, *81*(1), 112–136.

Cason, T. N., Friedman, D., & Wagener, F. (2005). The dynamics of price dispersion, or edgeworth variations. *Journal of Economic Dynamics and Control*, *29*(4), 801–822.

Claussen, J. C., & Traulsen, A. (2008). Cyclic dominance and biodiversity in well-mixed populations. *Physical review letters*, *100*(5), 058104.

Cook, R., Bird, G., Lünser, G., Huck, S., & Heyes, C. (2012). Automatic imitation in a strategic context: Players of rock-paper-scissors imitate opponents' gestures. *Proceedings of the Royal Society B: Biological Sciences*, *279*(1729), 780–786.

Cournot, A. (1838). *Recherches sur les principes mathematiques de la theorie des richesses*. Paris: Hachette. English translation (N. Bacon trans.): Research into the mathematical principles of the theory of wealth.

Danckert, J., Stöttinger, E., Quehl, N., & Anderson, B. (2012). Right hemisphere brain damage impairs strategy updating. *Cerebral Cortex*, *22*(12), 2745–2760.

Dyson, B. J. (2019). Behavioural isomorphism, cognitive economy and recursive thought in non-transitive game strategy. *Games*, *10*(3), 32.

Dyson, B. J., Steward, B. A., Meneghetti, T., & Forder, L. (2020). Behavioural and neural limits in competitive decision making: The roles of outcome, opponency and observation. *Biological psychology*, *149*, 107778.

Dyson, B. J., Sundvall, J., Forder, L., & Douglas, S. (2018). Failure generates impulsivity only when outcomes cannot be controlled. *Journal of Experimental Psychology: Human Perception and Performance*, *44*(10), 1483.

Dyson, B. J., Wilbiks, J. M. P., Sandhu, R., Papanicolaou, G., & Lintag, J. (2016). Negative outcomes evoke cyclic irrational decisions in rock, paper, scissors. *Scientific Reports (Nature Publisher Group)*, *6*(1), 20479.

Filipowicz, A., Anderson, B., & Danckert, J. (2016). Adapting to change: The role of the right hemisphere in mental model building and updating. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, *70*(3), 201.

Forder, L., & Dyson, B. J. (2016). Behavioural and neural modulation of win-stay but not lose-shift strategies as a function of outcome value in rock, paper, scissors. *Scientific Reports (Nature Publisher Group)*, *6*(1), 33809.

Frey, S., & Goldstone, R. L. (2013). Cyclic game dynamics driven by iterated reasoning. *Plos one*, *8*(2).

Gallagher, H. L., Jack, A. I., Roepstorff, A., & Frith, C. D. (2002). Imaging the intentional stance in a competitive game. *Neuroimage*, *16*(3), 814–821.

Garrido-da-Silva, L., & Castro, S. B. (2020). Cyclic dominance in a two-person rock-scissors-paper game. *International Journal of Game Theory*, 1–28.

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological review*, *103*(4), 650.

Hauert, C., Monte, S. D., Hofbauer, J., & Sigmund, K. (2002). Volunteering as red queen mechanism for cooperation in public goods games. *Science*, *296*(5570), 1129–1132.

Hegan, K. (2004). Hand to hand combat. *Rolling Stone.*

Hoffman, M., Suetens, S., Nowak, M. A., & Gneezy, U. (2012). An experimental test of nash equilibrium versus evolutionary stability. *Proc. Fourth World Congress of the Game Theory Society*, *145*.

Hopkins, E., & Seymour, R. M. (2002). The stability of price dispersion under seller and consumer learning. *International Economic Review*, *43*(4), 1157–1190.

Hu, W., Zhang, G., Tian, H., & Wang, Z. (2019). Chaotic dynamics in asymmetric rock-paper-scissors games. *IEEE Access*, *7*, 175614–175621.

Kalisch, G., Milnor, J., Nash, J., & Nering, E. (1954). *Some experimental n-person games.*

Kangas, B. D., Berry, M. S., Cassidy, R. N., Dallery, J., Vaidya, M., & Hackenberg, T. D. (2009). Concurrent performance in a three-alternative choice situation: Response allocation in a rock/paper/scissors game. *Behavioural Processes*, *82*(2), 164–172.

Kerr, B., Riley, M. A., Feldman, M. W., & Bohannan, B. J. (2002). Local dispersal promotes biodiversity in a real-life game of rock-paper-scissors. *Nature*, *418*(6894), 171–174.

Kirkup, B. C., & Riley, M. A. (2004). Antibiotic-mediated antagonism leads to a bacterial game of rock-paper-scissors in vivo. *Nature*, *428*(6981), 412–414.

Lach, S. (2002). Existence and persistence of price dispersion: An empirical analysis. *Review of economics and statistics*, *84*(3), 433–444.

Lakhar, R. (2011). The dynamic instability of dispersed price equilibria. *Journal of Economic Theory*, *146*(5), 1796–1827.

Lie, C., Baxter, J., & Alsop, B. (2013). The effect of opponent type on human performance in a three-alternative choice task. *Behavioural Processes*, *99*, 87–94.

Liptak, A. (2006). Lawyers won't end squabble, so judge turns to child's play. *The New York Times.*

Lopes, L. L. (1982). Doing the impossible: A note on induction and the experience of randomness. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *8*(6), 626.

Lopes, L. L., & Oden, G. C. (1987). Distinguishing between random and nonrandom events. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *13*(3), 392.

Morgenstern, O., & Neumann, J. V. (1953). *Theory of games and economic behavior*. Princeton university press.

Nash, J. F. (1950). Equilibrium points in n-person games. *Proceedings of the national academy of sciences*, *36*(1), 48–49.

Noel, M. D. (2007). Edgeworth price cycles: Evidence from the toronto retail gasoline market. *The Journal of Industrial Economics*, *5*(1), 69–92.

Nowak, M., & Sigmund, K. (1993). A strategy of win-stay, lose-shift that outperforms tit-for-tat in the prisoner's dilemma game. *Nature*, *364*(6432), 56–58.

Nowak, M., & Sigmund, K. (1990). The evolution of stochastic strategies in the prisoner's dilemma. *Acta Applicandae Mathematicae*, *20*(3), 247–265.

Nowak, M. A., & Sigmund, K. (1992). Tit for tat in heterogeneous populations. *Nature*, *355*(6357), 250–253.

Nowak, M. A., & Sigmund, K. (2004). Evolutionary dynamics of biological games. *science*, *303*(5659), 793–799.

O'Neill, B. (1987). Nonmetric test of the minimax theory of two-person zerosum games. *Proceedings of the national academy of sciences*, *84*(7), 2106–2109.

Posch, M. (1999). Win-stay, lose-shift strategies for repeated games—memory length, aspiration levels and noise. *Journal of theoretical biology*, *198*(2), 183–195.

Rapoport, A., & Budescu, D. V. (1992). Generation of random series in two-person strictly competitive games. *Journal of Experimental Psychology: General*, *121*(3), 352.

Rapoport, A., & Budescu, D. V. (1997). Randomization in individual choice behavior. *Psychological Review*, *104*(3), 603–617.

Schelling, T. C. (1958). The strategy of conflict. prospectus for a reorientation of game theory. *Journal of Conflict Resolution*, *2*(3), 203–264.

Schelling, T. C. (1960). *The strategy of conflict*. Harvard University Press.

Semmann, D., Krambeck, H. J., & Milinski, M. (2003). Volunteering leads to rock-paper-scissors dynamics in a public goods game. *Nature*, *425*(6956), 390–393.

Sepahvand, N. M., Stöttinger, E., Danckert, J., & Anderson, B. (2014). Sequential decisions: A computational comparison of observational and reinforcement accounts. *PloS one*, *9*(4).

Shannon, C. E. (1948). A mathematical theory of communication. *Bell system technical journal*, *27*(3), 379–423.

Sinervo, B., & Lively, C. M. (1996). The rock-paper-scissors game and the evolution of alternative male strategies. *Nature*, *380*(6571), 240–243.

Stöttinger, E., Filipowicz, A., Danckert, J., & Anderson, B. (2014). The effects of prior learned strategies on updating an opponent's strategy in the rock, paper, scissors game. *Cognitive Science*, *38*(7), 1482–1492.

Stöttinger, E., Filipowicz, A., Marandi, E., Quehl, N., Danckert, J., & Anderson, B. (2014). Statistical and perceptual updating: Correlated impairments in right brain injury. *Experimental brain research*, *232*(6), 1971–1987.

Szolnoki, A., Mobilia, M., Jiang, L. L., Szczesny, B., Rucklidge, A. M., & Perc, M. (2014). Cyclic dominance in evolutionary games: A review. *Journal of the Royal Society Interface*, *11*(100), 20140735.

Toupo, D. F., & Strogatz, S. H. (2015). Nonlinear dynamics of the rock-paper-scissors game with mutations. *Physical Review E*, *91*(5), 052907.

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological bulletin*, *76*(2), 105.

Vogel, C. (2005). Rock, paper, payoff: Child's play wins auction house an art sale. *The New York Times*.

Wang, Z., & Xu, B. (2014). Incentive and stability in the rock-paper-scissors game: An experimental investigation. *arXiv preprint arXiv:1407.1170*.

Wang, Z., Xu, B., & Zhou, H. J. (2014). Social cycling and conditional responses in the rock-paper-scissors game. *Scientific reports*, *4*, 5830.

West, R. L., & Lebiere, C. (2001). Simple games as dynamic, coupled systems: Randomness and other emergent properties. *Cognitive Systems Research*, *1*(4), 221–239.

Xu, B., Zhou, H. J., & Wang, Z. (2013). Cycle frequency in standard rock-paper-scissors games: Evidence from experimental economics. *Physica A: Statistical Mechanics and its Applications*, *392*(20), 4997–5005.

Yang, Q., Rogers, T., & Dawes, J. H. (2017). Demographic noise slows down cycles of dominance. *Journal of theoretical biology*, *432*, 157–168.

Zhang, R., Clark, A. G., & Fiumera, A. C. (2013). Natural genetic variation in male reproductive genes contributes to nontransitivity of sperm competitive ability in drosophila melanogaster. *Molecular ecology*, *22*(5), 1400–1415.

Zhou, H. J. (2016). The rock-paper-scissors game. *Contemporary Physics*, *57*(2), 151–163.

Chapter 2

Repeated rock, paper, scissors play reveals limits in adaptive sequential behavior

INTERIM SUMMARY

In the previous chapter, I argued that *mixed strategy equilibrium* (MSE) games offer a unique lens with which to study people's *opponent modeling* abilities over many repeated interactions. MSE games like rock, paper, scissors have a Nash Equilibrium (Nash, 1950) strategy of random play because any non-random dependency in one's moves can be exploited by a rational opponent. As a consequence, people playing a potentially fallible opponent are incentivized to look for and exploit such patterns. I argued that rock, paper, scissors is an ideal (and novel) environment for studying this exact ability because the space of possible patterns a player might exhibit or exploit in their opponent can be clearly and precisely spelled out, and people's ability to detect these patterns identified through their behavior and game outcomes. I first presented an overview of this structure which extends beyond any discussed in prior literature. Next, I analyzed data previously published in Brockbank and Vul (2020) indicating that people *do* empirically show evidence of exploiting opponents over many rounds in human dyads; I applied novel methods outlined in this work to investigate which patterns people exhibit and which ones their opponents likely take advantage of.

Having established that people can in fact build simple *behaviorist* models of their opponents in this task, the natural question is *how*? What patterns are they relying on? In essence, *what is the structure of behaviorist mental models in this setting*? The dyad experiment analyzed in the previous chapter, and the methods outlined there, provide only an approximate answer to these questions because dyad play is inherently adaptive. In chapter 2, I engage seriously with the question of just what it is people could be modeling in their opponent's moves, and which patterns they can adaptively modify in their own moves. I present results from two studies in which people play against algorithmic bot opponents that exhibit stable patterns in their moves or try to exploit patterns in human participants' own moves. This enables fine-grained control over the set of sequential dependencies that people are presented with and that their opponents adapt to, allowing for a precise specification of the complexity and structure of people's *behaviorist* opponent models. Results suggest that people can flexibly adapt to simple transition patterns in an opponent's moves and modify such patterns in their own moves, but fail to represent more complex dependencies, placing behaviorist intuitive psychology at odds with the rich, cognitivist mental models of sequential behavior explored in chapter 3.

Abstract

How do people adapt to others in adversarial settings? Prior work has shown that people often violate rational models of adversarial decision-making in repeated games. In particular, in *mixed strategy equilibrium* (MSE) games, where optimal action selection entails choosing moves randomly, people often do not play randomly, but instead try to *outwit* their opponents. However, little is known about the adaptive reasoning that underlies these deviations from random behavior. Here, we examine strategic decision-making across repeated rounds of rock, paper, scissors, a well-known MSE game. In experiment 1, participants were paired with bot opponents that exhibited distinct stable move patterns, allowing us to identify the bounds of the complexity of opponent behavior that people can detect and adapt to. In experiment 2, bot opponents instead exploited stable patterns in the human participants' moves, providing a symmetrical, matching bound on the complexity of patterns people can revise in their own behavior. Across both experiments, people exhibited a robust and flexible attention to *transition patterns* from one move to the next, exploiting these patterns in opponents and modifying them strategically in their own moves. However, their adaptive reasoning showed strong limitations with respect to more sophisticated patterns. Together, results provide a precise and consistent account of the surprisingly limited representational complexity of people's adaptive decision-making in this setting.

Keywords: adaptive reasoning, adversarial reasoning, opponent modeling, rock-paper-scissors

2.1 Introduction

People's ability to reason strategically and adapt to others in adversarial interactions lies at the heart of sports and games and is a hallmark of entertainment and storytelling; at a larger scale, it is crucial to negotiations and international relations. In these settings, examples of people's creativity, flexibility, and strategic sophistication abound. For instance, tennis star Andre Agassi famously beat world-class opponent Boris Becker by recognizing that every time he served the ball, Becker unknowingly stuck his tongue out in the direction he was about to serve.[1] On the other hand, our adversarial reasoning is constrained by challenges like remembering previous decisions (Rapoport & Budescu, 1997), recursive reasoning about others (Moulin, 1986), or merely searching a large space of potential actions. These limitations have allowed artificial intelligence systems to beat human competitors in a wide range of adversarial games, even those once thought to be far beyond the reach of strategic algorithms (Silver et al., 2016). *What kind of reasoning processes do people rely on in repeated adversarial interactions with others? And how does their behavior in such interactions reflect basic cognitive constraints?*

These questions have primarily been informed by prior work in game theory and behavioral economics, which uses decision-making in repeated games to formally describe how people balance risk and reward in adversarial and cooperative interactions (see, e.g., C. F. Camerer (2011)). For example, perhaps the most well-known application of game theory to understanding human behavior is the use of iterated choices in the Prisoner's Dilemma to describe the emergence of cooperation and reciprocity (Axelrod, 1984; Rapoport & Chammah, 1970). However, decision-making in repeated games has formed the basis for exploring an array of additional behaviors central to human intelligence and social inference, from recursive *theory of mind* (Moulin, 1986) to trust (C. Camerer & Weigelt, 1988). Adversarial and cooperative reasoning is frequently studied using repeated

[1] https://www.facebook.com/watch/?v=1249137535168463 - January 18, 2017

games in part because the existence of Nash Equilibrium (Nash, 1950) solutions provides a *rational model* of behavior in these settings, subject to particular assumptions about the players involved. In this way, the Nash Equilibrium is not just a mathematical formalism but offers a benchmark for understanding human decision-making; people's choices across repeated interactions are often characterized in terms of their departures from optimal reasoning or from the underlying assumptions of equilibrium play (C. F. Camerer, 2011).

To illustrate, in *mixed strategy equilibrium* (MSE) games like rock, paper, scissors, or matching pennies, people's behavior over repeated interactions often fails to reflect the Nash Equilibrium. Nash Equilibrium play in MSE games requires that players choose their moves *randomly*; any non-random dependency in their behavior (e.g., a bias toward *rock*) is exploitable by a rational opponent. However, a large body of work on *subjective randomness* suggests that people exhibit poor performance when asked to either generate random sequences or detect them (Bar-Hillel & Wagenaar, 1991). Instead, people typically rely on simple biases like an over-representation of alternations relative to repeats (Lopes & Oden, 1987; Tversky & Kahneman, 1972). More recent work has proposed that perceptions of randomness may in fact reflect rational statistical inference about the absence of detectable patterns (Griffiths et al., 2018). However, biased judgments about what constitutes a random sequence are nonetheless at odds with Nash Equilibrium behavior in MSE games; a savvy adaptive player can *outwit* an opponent who relies on such biases when selecting their moves. Indeed, prior work exploring people's behavior in repeated MSE games found that while adversarial settings may improve people's ability to produce unpredictable sequences of actions, the same underlying biases about randomness persist (Budescu & Rapoport, 1994; Rapoport & Budescu, 1992).[2] These biases are so ingrained that patterns associated with subjective randomness arise even in decisions by professional athletes, who are highly incentivized to avoid such predictability (Palacios-Huerta, 2003; Walker & Wooders, 2001). Thus, MSE games represent a setting in which adversarial reasoning veers sharply and predictably from the Nash Equilibrium even when people do their best not to.

How then can we explain people's non-random adversarial behavior in mixed strategy equilibrium games? Consistent with the broader literature on subjective perceptions of randomness, early accounts of decision-making in MSE games largely focused on the kinds of iterated reasoning about one's own moves that might produce *seemingly* random behavior (Rapoport & Budescu, 1997). More recent work has emphasized the use of generic *heuristics* such as "win-stay, lose-shift," which offer simple decision procedures and may reflect intuitive responses to losses and gains (Dyson et al., 2018). For example, people show evidence of win-stay, lose-shift behavior in repeated rock, paper, scissors games against bot opponents that choose their moves randomly (Dyson et al., 2016; Forder & Dyson, 2016), perhaps reflecting the intuitive response to an unexploitable adversary; recent work has also found that such heuristic responding is prevalent against shuffled human opponents (Wang et al., 2014) and stable human opponents (Brockbank & Vul, 2020). Given the role that such heuristics play in other forms of learning and inference (Bonawitz et al., 2014; Gigerenzer & Goldstein, 1996), these simple, stable strategies likely form an important part of people's adversarial behavior, especially against an unexploitable opponent.

However, accounts of adversarial behavior that are based on stable heuristics limit the role of *adaptive* processes that underlie much of our strategic adversarial reasoning (recall, for example, Andre Agassi's insight about Boris Becker described above). In other words, it is clear that people playing MSE games do not implement a Nash Equilibrium strategy when choosing moves, but accounts of this deviation that invoke heuristics alone may overlook the learning processes and online reasoning about an opponent that people can recruit in this setting (Brockbank & Vul, 2021). In fact, prior work suggests that flexible and adaptive reasoning forms an important part of people's responding in repeated

[2] The persistence of people's subjective randomness judgments led Rapoport and Budescu (1992) to conclude: "Cognitive psychology has engendered few examples of so much support for and agreement among researchers about the prevalence of a cognitive bias" (p. 352).

MSE games. When paired with rock, paper, scissors opponents that favor a particular move, people typically learn to exploit them so long as the bias is sufficiently strong (Kangas et al., 2009; Lie et al., 2013). More recent work has suggested that people can also exploit opponents that exhibit more complex patterns in their move choices contingent on prior moves (Dyson et al., 2020; Dyson et al., 2018). But how much does this behavior reflect ongoing adaptive reasoning? Stöttinger et al. (2014) found that when paired with multiple bot opponents that displayed distinct biases towards particular moves or transitions between moves, people successfully exploited these patterns from one opponent to the next, though their ability to do so was sensitive to what kind of opponent they had previously encountered. The ability to outwit an opponent whose moves are based on novel but predictable underlying patterns requires flexible learning processes to detect and exploit such patterns. And this adaptive reasoning appears to extend to more fluid and dynamic interactions with human opponents as well. In Brockbank and Vul (2020), participants played 300 rounds of rock, paper, scissors against another human participant. Among the dyads, one player was able to consistently beat their opponent significantly more often than would be expected by chance. Taken together, these results suggest that in repeated MSE games like rock, paper, scissors, people exhibit *adaptive* adversarial reasoning about their opponent. However, these results lack a systematic account of people's adaptive behavior in this setting: Which patterns in another player's actions can they successfully learn and which ones are out of reach? And how well can people avoid being similarly exploited?

The finding that people's behavior in mixed strategy equilibrium games recruits *adaptive* reasoning about their opponent is consistent with many other adversarial interactions which may involve flexible, on-the-fly decision-making (e.g., games like chess). *What, then, can behavior observed in MSE games like rock, paper, scissors tell us about how people adapt to an adversarial opponent?* Critically, RPS offers little opportunity for expertise, thus differentiating it from games like chess or tennis. This means that in a laboratory setting, RPS can be used to explore people's adaptive, adversarial reasoning abilities independent of prior experience. Instead, the game's simple rules and small set of available actions present participants with a seemingly trivial challenge: Over many repeated interactions, beating one's RPS opponent is a matter of detecting exploitable patterns in that person's moves while minimizing such exploitability in one's own choices. The scope of the behavioral patterns or dependencies that people can reliably detect in an opponent's moves and in their own can therefore inform our understanding of the representational flexibility of people's adaptive reasoning. Finally, because of the relatively simple structure of the game, the sequential patterns in a player's moves can be carefully measured and compared (Brockbank & Vul, 2021; Dyson, 2019). In this way, we not only ask what patterns people can adapt to in an opponent or minimize in their own actions, but what kind of underlying formal structure is shared across these patterns.

The current work explores these questions using people's behavior in repeated rounds of rock, paper, scissors with an algorithmic "bot" opponent. In experiment 1, we pair participants with one of seven stable bots, each of which exhibits a different sequential dependency in its move choices. These dependencies vary in their underlying complexity, allowing us to precisely assess the degree to which people track and exploit different behavioral contingencies affecting an opponent's moves. We find that people are highly adaptive against opponents that exhibit simple *transition* patterns, but otherwise show minimal adaptation to more complex opponents. In experiment 2, we ask whether these same limits hold for detecting and revising exploitable patterns in one's *own* behavior. Participants were once again paired with a bot opponent, but this time each bot chose its moves by trying to exploit a unique pattern in the participant's moves. Here, we examine people's ability to avoid or counteract such exploitative behavior. Once again, we find that people are strongly adaptive against bots that track simple transition patterns in participant moves, but show little flexibility otherwise. Together, our results suggest that the *hypothesis space* of behavioral patterns people draw on in this setting to understand their opponent's moves or their own is limited, but that adaptive reasoning is flexible within these limits.

2.2 Experiment 1: Stable Bot Opponents

Experiment 1 pitted participants against predictable bot opponents that chose their moves by following one of seven increasingly complex sequential dependencies. If participants can reliably beat a bot opponent that exhibits a stable pattern in its moves, this suggests people are able to adapt to that particular dependency in an adversarial setting. We investigate the level of behavioral complexity that people can detect and exploit across repeated interactions with a bot opponent.

2.2.1 Participants

Participants were a convenience sample of 218 University of California, San Diego undergraduate students who received course credit for their participation. One student's data was removed due to technical issues during data collection which prevented completion of the experiment. Our sample size was chosen to have a minimum of 30 participants in each condition (i.e., against each bot opponent). This gave us 90% power to detect an effect size of $d = 0.61$ in our estimate of participant win rates against each bot; under a conservative assumption of uniformly distributed win rates, this effect size amounts to an average win rate of approximately 51%. Informed consent was obtained from all participants in accordance with the Institutional Review Board's approved protocol. Participants completed the experiment in a web browser online.[3]

2.2.2 Task Overview

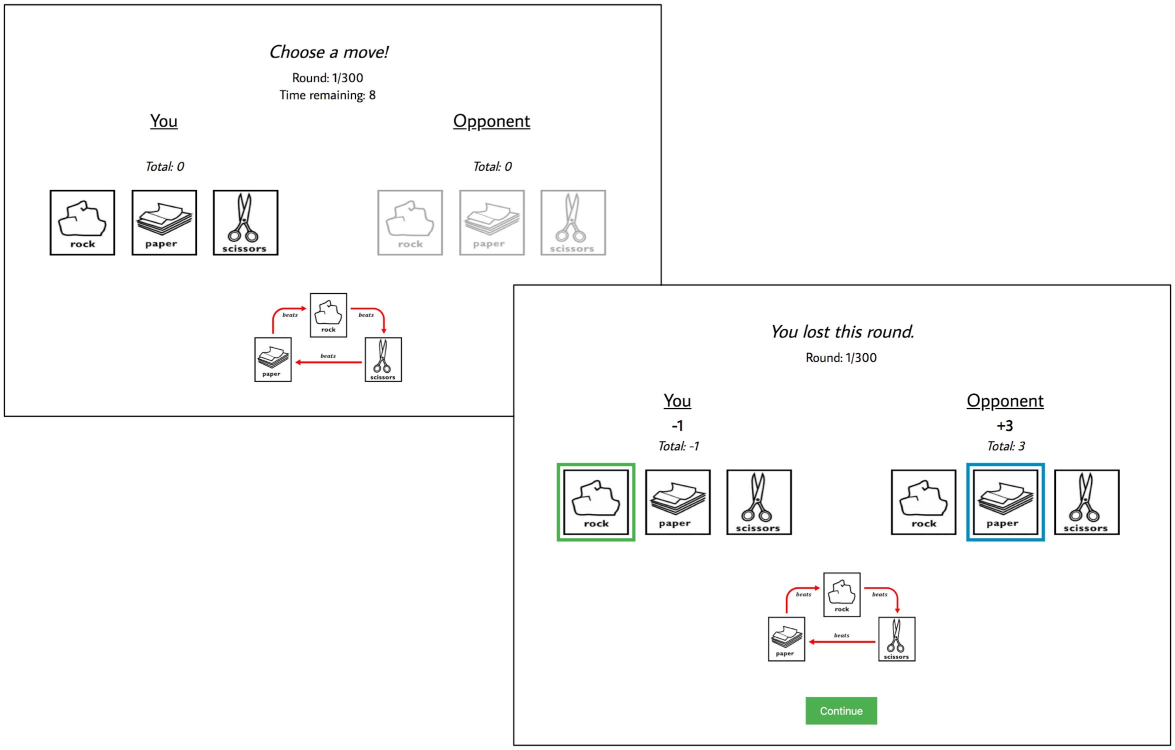
Participants began by clicking through a set of instructions introducing the game of rock, paper, scissors and noting that they would be playing 300 rounds against a fixed

[3] The data and code for this experiment and for all analyses reported here can be found on github at: https://github.com/erik-brockbank/rps-bot-manuscript-public.

opponent (they were told it would be the same opponent the whole time but were not told anything more about the opponent's identity). Upon completion of the instructions, participants were randomly assigned to one of seven bot opponent conditions, described in detail below. In each round, they were shown a set of clickable "cards" with rock, paper, and scissors icons and instructed to choose a move (Figure 2.1). They were given 10 seconds to choose their move each round. Once a participant had chosen a move, they could not change their selection. After selecting a move, participants were shown a results screen indicating their own move, their opponent's choice, the results of the round, and the points each player received for that round (see Figure 2.1). Participants were given 3 points for a win, 0 points for a tie, and -1 points for a loss. These values were chosen to maintain engagement by giving participants greater opportunity for positive accumulation of points. However, this imbalanced allocation of points across outcomes does not impact the optimal play for participants; since there is a single move every round that would give them a high probability of winning (see below), playing the move that's expected to win is best under *any* setup in which wins offer more points than ties or losses. At the end of each round, participants could view the results of the round for as long as they wanted before clicking a button to proceed to the next round.

Throughout the game, participants were shown a graphic illustrating each move's relation to the others, a tally of rounds completed towards the total, and the cumulative points each player had accrued so far (see Figure 2.1). The displayed scores and rounds completed served to motivate participants to best their opponent over the 300 rounds, while the graphic of each card's relationship to the others minimized the risk of them misunderstanding or forgetting the rules. Most participants completed the RPS game in under 30 minutes (mean time from round one to round 300: 691s, SD: 259s).

Following completion of the game, participants completed a brief post-experiment questionnaire with a free response prompt asking about how they had chosen their moves (*"Please describe any strategies you used to try and beat your opponent"*) and a set of five slider scale questions further assessing their strategic decision-making during the task. The first question probed participants' beliefs about their opponent (*"My opponent was a real person and not a robot"*), the next three probed participants' motivation and whether they were watching for patterns in their opponent's moves, and the final question asked about their explicit awareness of any patterns in their opponent's move choices (*"There were noticeable patterns in my opponent's moves that allowed me to predict their next move"*). Response options ranged from 1 ('Strongly disagree') to 7 ('Strongly agree'). We do not analyze these responses here but the data is available at the repository linked above.



Choose a move!

Round: 1/300

Time remaining: 8

You lost this round.

Round: 1/300

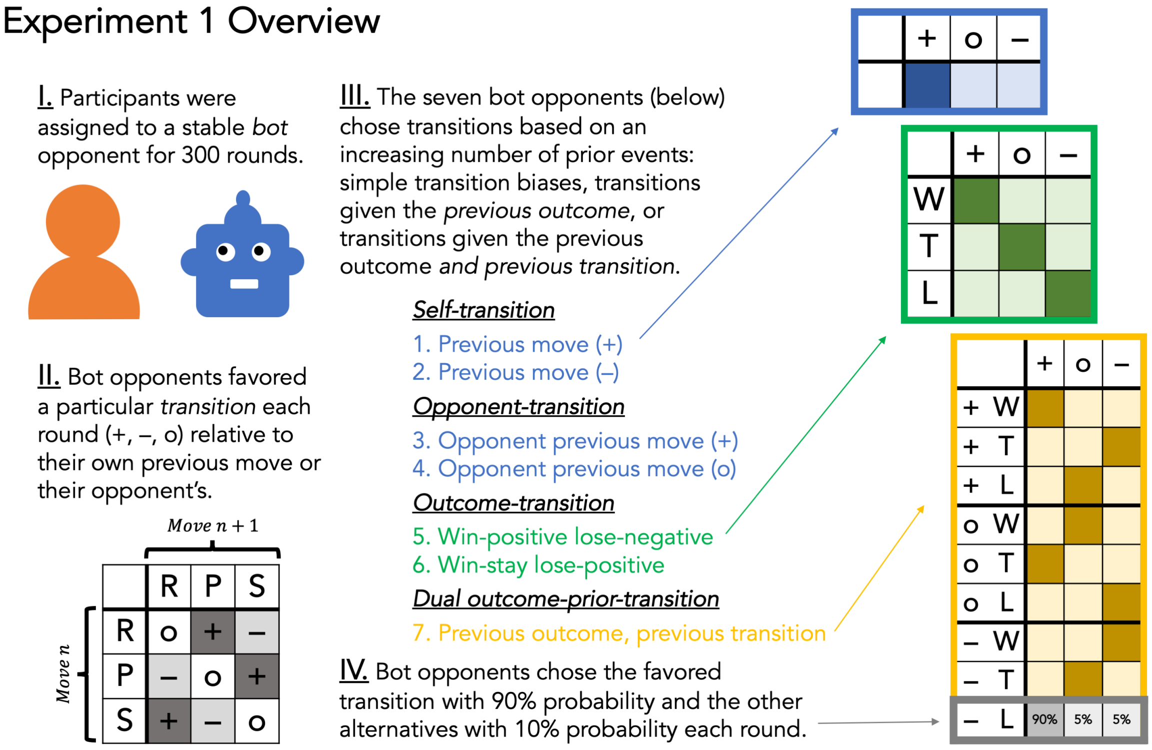
Figure 2.1. The stages of each rock, paper, scissors round. Top left: participants had 10s to select their move by clicking one of the three "cards." Bottom right: participants were shown the results of the round, along with updated points for each player, before clicking to proceed to the next round.

2.2.3 Bot Opponent Behavior

Participants were paired with one of seven bot opponents that exhibited a distinct sequential pattern or "strategy" in its move choices. Concretely, each bot's strategy dictated a particular move each round based on events from the preceding round(s). Bots chose this move 90% of the time and each of the other (non-strategic) moves 5% of the time to allow for some noise in their strategy. The differing complexity of the bot strategies can be described in terms of the previous events that dictated their move choice each round; in this way, the bot strategies form a hierarchy with an increasing number of events affecting their move choices. These sequential dependencies in the bots' decisions were chosen to encapsulate the full range of patterns that we expected human players to be able to detect and exploit in an opponent. The patterns have been formally described in prior work (Brockbank & Vul, 2021; Dyson, 2019) and have been shown to arise in repeated games of rock, paper, scissors among human dyads (Brockbank & Vul, 2020), suggesting that people may be sensitive to these patterns in strategic adversarial settings. We walk through this "strategy space" below (see overview in Figure 2.2).

Transition dependencies

The simplest level at which the bots exhibited dependencies in their move choices was based on a single previous move. This *transition dependency* can be expressed in three possible ways. A "positive" or "upward" transition (+) bias amounts to favoring a move that would *beat* the previous move (e.g., playing "paper" most often following "rock"). A "negative" or "downward" transition (−) bias involves instead choosing the move that would *lose* to the previous move, for example transitioning most often from "rock" to "scissors", "scissors" to "paper", and "paper" to "rock". Finally, a "stay" transition (0) bias is one in which a bot is most likely to repeat the same move again. Critically, such transition biases can be exhibited relative to one's own previous move *or an opponent's previous move*. Thus, a bot that repeats moves more often than either alternation (+/−)



Experiment 1 Overview

I. Participants were assigned to a stable *bot* opponent for 300 rounds.

II. Bot opponents favored a particular *transition* each round (+, –, o) relative to their own previous move or their opponent's.

III. The seven bot opponents (below) chose transitions based on an increasing number of prior events: simple transition biases, transitions given the *previous outcome*, or transitions given the previous outcome *and previous transition*.

Self-transition

1. Previous move (+)
2. Previous move (–)

Opponent-transition

3. Opponent previous move (+)
4. Opponent previous move (o)

Outcome-transition

5. Win-positive lose-negative
6. Win-stay lose-positive

Dual outcome-prior-transition

7. Previous outcome, previous transition

IV. Bot opponents chose the favored transition with 90% probability and the other alternatives with 10% probability each round.

Figure 2.2. Experiment 1 overview.

can be described has having a "stay" (0) bias in its *self-transitions* and a bot that most often plays what its opponent played in the previous round can be described as having a "stay" (0) bias in its *opponent-transitions*.

Outcome-transition dependencies

The transition dependencies above can be expanded by including prior events or *contingencies* that affect a player's transition bias (see Figure 2.2). The simplest such extension involves a bias towards different transitions depending on the previous game's *outcome*. For example, a player might prefer to repeat the same move when it is successful (after a *win*) but transition up (+) or down (–) after a tie or loss. Biases towards *win-stay, lose-shift* responding described in human play against random or shuffled opponents (Dyson et al., 2016; Wang et al., 2014) falls under this level of dependency. Critically, while a simple transition bias can be described as occupying one of three possible states (favoring *up*, *down*, or *stay* relative to a previous move as described above), outcome-dependent transition patterns involve biases across nine contingent states: three possible transitions (+, −, 0) for each previous outcome (*win*, *loss*, or *tie*; see Figure 2.2). It is worth noting that because an outcome is a function of the player's previous move and their opponent's, outcome-dependent transitions can be equivalently stated in terms of *self-transitions* and *opponent-transitions* (e.g., a positive (+) *self-transition* after a win is equivalent to a negative (−) *opponent-transition* after a win). We describe all bot outcome-transition dependencies below in terms of their self-transition biases after each outcome.

Dual outcome-prior-transition dependencies

In addition to the previous round's outcome, a player's transition from one round to the next might show a further dependency on their *previous transition.* For example, if they transitioned *up* (+), *down* (−), or *stay* (0) in the previous round and *won*, this might lead them to make the same transition again in the next round (i.e., transitioning *up* and then winning might make another *up* transition more likely). Here, the bias towards a particular *self-transition* in each round is contingent on nine unique previous event combinations (a previous *up* transition and a *win*, a previous *up* transition and a *tie*, ...), leading to 27 different directions that transitions may be biased based on these prior events (see Figure 2.2). Dependencies described at this level are therefore significantly more complex than the simple transition biases outlined above.

Although further expansions of RPS move dependencies are possible, the patterns outlined here represent a principled approach to capturing the full range of structured behavior that people might adapt to in adversarial interactions: an opponent's previous move (*transition dependencies*), the combination of both players' previous moves (*outcome-transition dependencies*), and a further inclusion of the opponent's choice *two* moves prior (*dual outcome-prior-transition dependencies*). Below we describe how the bot opponents in the present experiment spanned these behavioral complexity levels.

Bot strategies

The seven bot opponents that participants faced in the current experiment were chosen to encapsulate the full range of behavioral dependencies described above. The bot "strategies" are illustrated in Figure 2.2 and outlined below in increasing order of complexity.

Transition strategies:

- **Previous move** (+) bot most often chose the move that constituted a positive or upward transition relative to its own previous move, i.e., the move that would beat what it had previously played.

- **Previous move** (−) bot favored the move each round that constituted a negative or downward transition relative to its own previous move, i.e., the move that would lose to what it had previously played.

- **Opponent previous move** (+) bot favored the move each round that constituted a positive transition *relative to its opponent's previous move*. In other words, it typically chose the move that would beat what its opponent had just played.

- **Opponent previous move** (0) bot primarily chose the move each round that constituted a "stay" transition, once again relative to its *opponent's* previous move, essentially copying its human opponent from one round to the next.

Outcome-transition strategies:

- **Win-stay lose-positive** bot favored stay self-transitions (0) after a win, positive self-transitions (+) after a loss, and negative self-transitions (−) after a tie.

- **Win-positive lose-negative** bot made positive self-transitions (+) after a win, negative self-transitions (−) after a loss, and stay self-transitions (0) after a tie.

Dual outcome-prior-transition strategies:

- **Previous outcome, previous transition** bot favored distinct self-transitions based on each unique *combination* of its previous self-transition *and* the previous round outcome that had resulted from its previous self-transition.

The seven bot strategies in this experiment represent a strategic sampling of all the possible *transition* dependencies, *outcome-transition* dependencies, and *dual outcome-prior-transition* dependencies described above that can be enumerated for a given player. Though many such biases are possible within each of these categories, we have chosen only those which satisfy two additional criteria. First, the bot strategies above cannot be described at a simpler complexity level (for example, an *outcome-transition* agent that transitions *up* most often after *every* previous outcome is no different from a simpler *transition* agent; see Dyson (2019) for further discussion of this). Second, we have excluded several potential bot strategies that could be exploited in simple ways without sensitivity to the underlying pattern in the bot's behavior. For example, among the *opponent-transition* dependency bots described above (`Opponent previous move (+)` and `Opponent previous move (0)`), a third strategy not included here involves favoring downward transitions (−) relative to an opponent's previous move each round. However, pitted against such a strategy, a player can exploit their opponent by simply playing the same move over and over. In this case, it might be argued that a human paired with a bot implementing this strategy did not so much adapt to the bot's dependence on the human's previous move as discover the success of playing the same move repeatedly without any regard for the bot's moves. In other words, exploiting the bots above requires an adaptive response that extends beyond merely playing the same move over and over (or beating a bot that does the same). The seven bot strategies in this experiment therefore include both a broad sampling of the patterns people might adapt to, as well as a selective or *diagnostic* set of patterns for understanding people's ability to adapt to their bot opponent.

2.2.4 Measuring adaptive reasoning

To understand how sensitive people were to stable patterns in their opponent's behavior, we measured participants' *win percentage* against each bot opponent. If participants fail to adapt to the sequential pattern exhibited by their bot opponent, we expect them to perform at chance, winning 1/3 of rounds in a given block of trials. Win rates greater than 33% indicate successful exploitation by participants. Together, we compare participants' win percentage against each of the seven bot opponents to better understand the level of behavioral complexity that people can adapt to over repeated interactions with their opponent. Which bot strategies do participants successfully exploit over 300 rounds and which ones do they fail to adapt to?

2.2.5 Results

Adaptation to bot strategies

How well did people detect and adapt to regularities of varying complexity when playing bot opponents? Figure 2.3 (Left) shows *overall* participant win percentages against each bot strategy across the 300 rounds. Participants were highly successful at exploiting the four simple *transition* bots (`Previous move (+)`, `Previous move (−)`, `Opponent previous move (+)`, and `Opponent previous move (0)`). Average win percentages ranged from 59.1% (*SE* = 3.28%) to 66.5% (*SE* = 3.01%) against these opponents. Not surprisingly, participant win percentages were significantly higher than chance in these conditions (`Previous move (+)`: $t(31) = 11.05$, p<.001; `Previous move (−)`: $t(29) = 8.99$, p<.001; `Opponent previous move (+)`: $t(30) = 7.84$, p<.001; `Opponent previous move (0)`: $t(31) = 9.97$, p<.001). In contrast, participants showed only moderate success against the *outcome-transition* strategies (`Win-stay lose-positive` and `Win-positive lose-negative`). Average win rates against these strategies were 41.3% (*SE* = 2.07%) and 39.7% (*SE* = 2.09%), respectively. Though participants performed above chance overall in these conditions as well (`Win-stay lose-positive`: $t(29) = 3.82$, $p<.001$; `Win-positive lose-negative`: $t(30) = 3.06$, $p = .005$), they were far less successful than those matched with the *transition* bots. Finally, participants performed poorly against the most complex *dual outcome-prior-transition* strategy (`Previous outcome, previous transition`); average win rate was 34.1% ($SE = 0.7\%$) and did not differ significantly from chance ($t(30) = 1.04$, $p = .31$). In this way, the seven strategies tested here decompose adaptive reasoning into patterns that can be detected and exploited *strongly*, *partially*, and *not at all*.

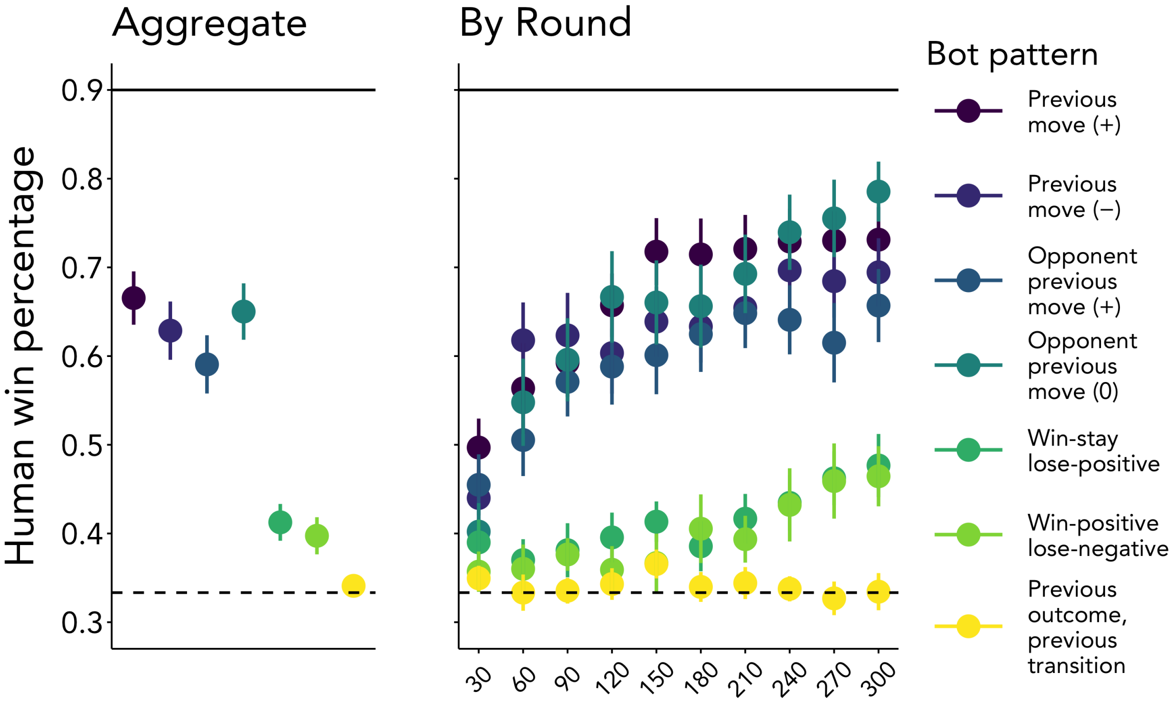


Figure 2.3. Performance against RPS bot strategies. (Left) Overall participant win percentage against each bot. (Right) Participant win percentage against each bot over the course of the experiment. Error bars reflect standard error of participant averages. Dashed lines indicate chance performance, while solid lines indicate optimal performance. Participants were highly successful at detecting and exploiting transition dependencies but showed little adaptation to more complex behavior patterns.

Participants' learning trajectories against each bot tell a similar story. Figure 2.3

(Right) shows win rates against each strategy over 10 sequential segments. For the simple *transition* bots (`Previous move (+)`, `Previous move (−)`, `Opponent previous move (+)`, and `Opponent previous move (0)`), participants rapidly detected the opponent's strategy and successfully exploited it for the majority of the experiment; learning rates and maximum performance are similar against all four *transition* dependencies. Meanwhile, people's ability to exploit the *outcome-transition* bots (`Win-stay lose-positive` and `Win-positive lose-negative`) arose only in the last 100 rounds of the experiment, and never reached performance levels comparable to the *transition* strategies. This suggests that the *outcome-transition* strategies were only partially exploitable (we describe this further below). Finally, consistent with aggregate performance, participants never succeeded above chance against the *dual outcome-prior-transition* bot (`Previous outcome, previous transition`).

Decomposing adaptive behavior

What explains participants' failure to adapt to the *outcome-transition* dependencies as effectively as the simpler *transition* bots? Incomplete exploitation of a complex bot strategy could arise from selective learning of the dependency, wherein only part of the opponent's contingency structure is effectively used against them, or degraded overall learning such that participants exploit the full scope of the dependency, albeit noisily. These constitute distinct accounts of the learning process underlying people's adaptive reasoning in this setting. Figure 2.4 shows human win rates during the last 100 rounds against the *transition* and *outcome-transition* bots, separated by each of the individual dependencies that dictate the bot's move choice. This illustrates the degree to which participants exploited the *full dependency structure* of their bot opponent. Against the four simple *transition* strategies (Figure 2.4 A–D), win percentages were significantly greater than chance following all bot or participant previous moves (all ps<.001) and win rates following each previous move were not significantly different from each other for three

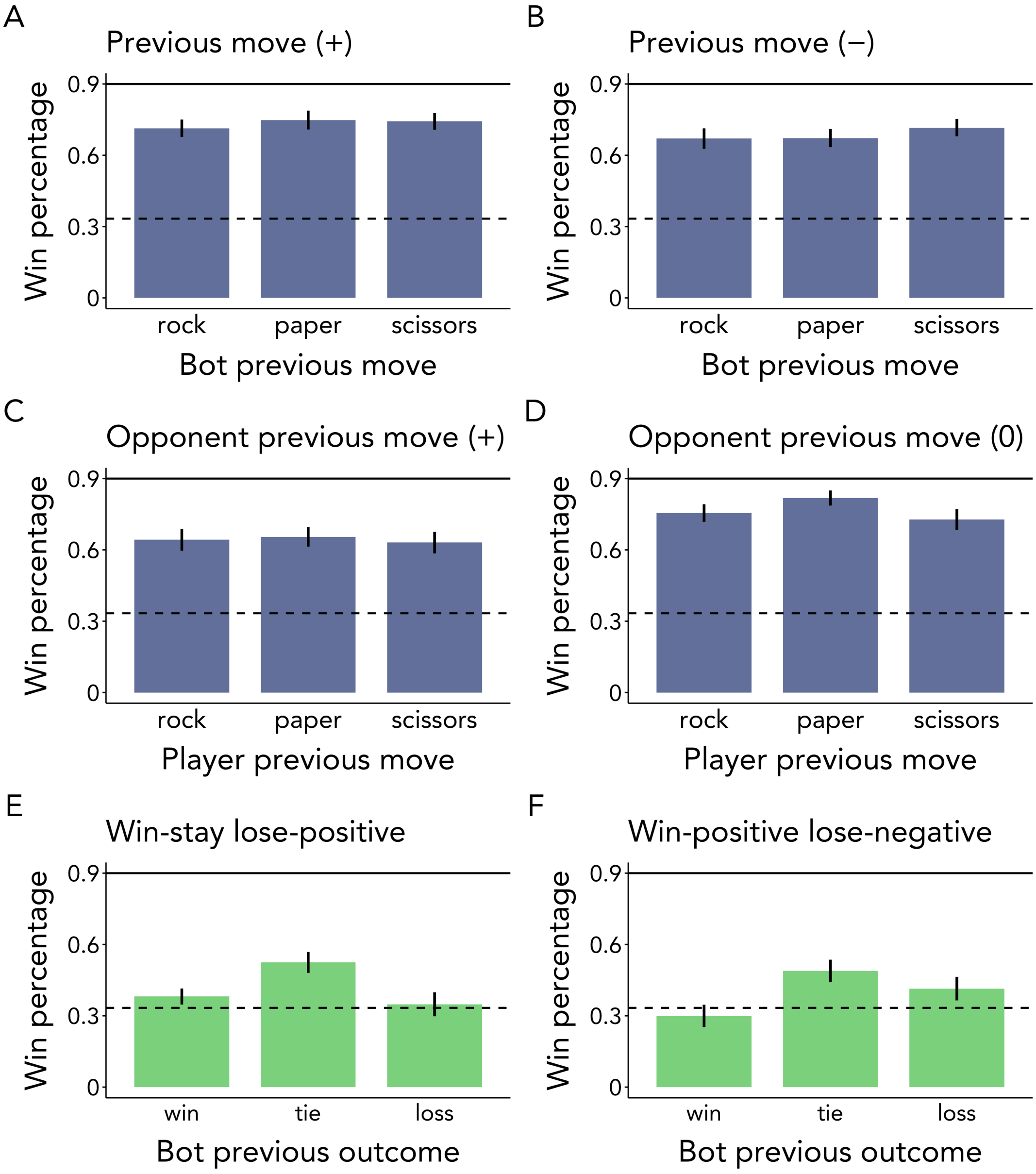


Figure 2.4. Participant win percentage against bot opponents following each of the prior events that dictate the bot's move choice. (A)-(D) Conditional win rates against the transition strategies based on each previous bot or participant move. (E)-(F) Conditional win rates against the outcome-transition strategies based on each previous bot outcome. Error bars reflect standard error of participant averages. Dashed lines indicate chance performance, while solid lines indicate optimal performance. People showed uniformly successful adaptation against the transition dependencies; partial adaptation to the outcome-transition dependencies was primarily based on exploiting individual contingencies in each bot's strategy (behavior after ties).

of the four *transition* opponents (`Previous move (+)`: $F(2) = 0.73$, $p = .49$; `Previous move (−)`: $F(2) = 1.53$, $p = .23$; `Opponent previous move (+)`: $F(2) = 0.19$, $p = .83$; `Opponent previous move (0)`: $F(2) = 5.10$, $p = .01$), suggesting a fairly uniform win rate in each condition. Thus, participants paired with the *transition* bots exploited the full contingency structure of their opponents' behavior.

However, against the two *outcome-transition* strategies (Figure 2.4 E–F), participants only won reliably after a *tie* outcome, indicating that they effectively learned *individual components* of the opponent's strategy. To illustrate, win percentages following each previous outcome varied significantly by outcome (`Win-stay lose-positive`: $F(2) = 3.77$, $p = .03$; `Win-positive lose-negative`: $F(2) = 3.29$, $p = .04$). These non-uniform win percentages were driven by the fact that for both of these strategies, win percentages differed significantly from chance following a tie (`Win-stay lose-positive`: $t(29) = 4.34$, $p<.001$; `Win-positive lose-negative`: $t(30) = 3.30$, $p = .003$) but *not* following a win or loss. This selective adaptation suggests that for strategies participants partially exploited, their dependency learning lacked the representational complexity to exploit the full scope of the strategy, instead exploiting specific behaviors *within* the more complex strategies.

2.2.6 Discussion

In this experiment, we explored the basis for people's *adaptive* adversarial reasoning abilities in a simple mixed strategy equilibrium game. Participants played 300 rounds of rock, paper, scissors against a bot opponent that exhibited one of seven distinct patterns in its move choices. These patterns increased in complexity from simple *transition* dependencies based on the bot's own or the participant's previous move, to intermediate *outcome-transition* dependencies where transition biases varied across prior outcomes (e.g., "win-stay, lose-shift"), to the most complex bot opponent, whose transitions from one round to the next were contingent on both the previous outcome and the bot's own previous transition. We examine how well participants exploited their bot opponent to better understand the complexity of people's adaptive reasoning in this setting.

Results contain two key findings. First, the seven sequential dependencies in bot moves provide a clear continuum of patterns that people can adapt to. Participants learned rapidly and were highly successful against the *transition* opponents, showed some success against the *outcome-transition* bots (primarily in the final third of the game), and performed at chance against the complex *dual outcome-prior-transition* strategy. Second, a close examination of participants' *conditional* win rates following each round outcome shows that people were not adapting to the two *outcome-transition* strategies uniformly. Instead, their partial success appeared largely isolated to exploiting individual transitions in the bot's strategy (i.e., those following a *tie*).

Broadly, these results suggest that participants were only sensitive to the simplest patterns in their opponent's behavior. The bots exhibiting *self-transition* dependencies based on their own prior move (`Previous move (+)`, `Previous move (−)`) were primarily cycling through moves from one round to the next ("rock", "paper", "scissors", "rock", ...). This was likely a salient pattern for participants. Participants' strong performance against the *opponent-transition* bots (`Opponent previous move (+)` and `Opponent previous move (0)`) suggests that the bot copying their previous move or choosing a move that would have beaten the participant's previous move may have been a clear pattern as well. In all of these cases, the complexity of the adaptive response was minor—optimal exploitation of these bots can be achieved by participants implementing particular self-transition biases in their own moves. However, participants showed little ability to adapt to opponents that exhibited additional complexity in their move strategies. Despite the fact that "win-stay, lose-shift" behavior has been observed in people's *own* move choices over repeated games against bot opponents (Dyson et al., 2016; Forder & Dyson, 2016), shuffled human opponents (Wang et al., 2014), and stable human opponents (Brockbank & Vul, 2020), participants adapted minimally to these dependencies in the *outcome-transition* bots. Taken together, results suggest that people have a limited *hypothesis space* for representing sequential structure in an opponent's moves, but will readily and flexibly exploit the simplest structures when they are present.

However, a central challenge for outwitting the bot opponents in the current task comes from the fact that participants must search for dependencies in their opponent's moves while also ostensibly avoiding predictable patterns in their *own* actions. One explanation for the seemingly small set of opponent behaviors that participants successfully exploited in the current experiment is that adaptive responses against stable bots are only half of the picture; the current experiment ignores the kinds of patterns people might be attending to in their own decisions. It may be that the representational complexity of this self-monitoring is far more expansive than the patterns that participants adapted to in their bot opponents' moves. *A complete account of adaptive adversarial reasoning should therefore consider people's ability to detect and modify sequential patterns in their own moves.* Experiment 2 addresses this second aspect of strategic decision-making.

2.3 Experiment 2: Adaptive Bot Opponents

In experiment 1, participants paired with stable bot opponents successfully adapted to dependencies in their opponent's move choices under simple conditions, but largely failed to exploit more complex sequential dependencies. However, it may be that people show a greater ability to represent complex regularities in their *own* behavior relative to their opponent's. In essence, sophisticated adaptive reasoning may be driven primarily by trying to generate or avoid predictable behavior patterns in one's own moves while responding to only the most glaring patterns exhibited by an opponent. The more complex behaviors displayed by the bots in experiment 1 arise in people's own move choices against human and bot opponents (Brockbank & Vul, 2020; Dyson et al., 2018; Dyson et al., 2016; Wang et al., 2014), suggesting that people are potentially vulnerable to exploitation of these patterns. In light of this, the current experiment tests whether people are capable of greater adaptive introspection about their own behavior and therefore able to temper or modify complex regularities in their actions, even though they fail to exploit these same patterns in others. Or, are the dependencies people can alter in their own moves just as limited as the ones they reliably exploit in an opponent?

We test these questions by once again evaluating people's behavior over 300 rounds of rock, paper, scissors against a bot opponent. However, rather than choosing their moves according to a fixed dependency structure as in experiment 1, the bots in the current experiment *adapted to the human participants*. Concretely, each bot tracked a distinct sequential regularity in its human opponent's moves over the course of the game and tried to exploit this regularity. On every round, the adaptive bot identified its human opponent's most likely move based on the particular pattern it was tracking, then chose its own move accordingly. Thus, for participants to succeed against their adaptive bot opponent required reducing the degree to which they exhibited the particular pattern the bot was exploiting. The patterns that the adaptive bots exploited in their opponents mirrored the structure of those exhibited by the stable bots in experiment 1 and varied in their underlying complexity. This allows us to precisely characterize the robustness of people's capacity to revise increasingly complex sequential patterns in their own decisions.

2.3.1 Participants

Participants were 194 undergraduate students who received course credit for their participation. One participant was removed because of technical error and a second was excluded due to clear evidence of not trying (i.e., choosing the same move to their own detriment in the vast majority of rounds). Participants were randomly assigned to one of eight adaptive bot conditions, described in further detail below. Our sample size was chosen to have a minimum of 20 participants in each bot condition. This gave us 90% power to detect an effect size of $d = 0.77$ in our estimate of bot win rates against each participant (or, symmetrically, participant win rates against each bot), an average win rate of roughly 55% assuming uniformly distributed win rates. Informed consent was obtained from all participants in accordance with the Institutional Review Board's approved protocol. Participants completed the experiment in a web browser online.[4]

2.3.2 Task Overview

The procedure for the experiment was identical to experiment 1 (Figure 2.1) with the notable exception that participants were now paired with an *adaptive* bot opponent. This opponent chose its moves in an effort to maximize expected win probability in each round, using the participant's decisions on prior rounds to predict their next move (see Figure 2.5). As before, participants were not told anything about their opponent's identity. All additional aspects of the experiment were identical to experiment 1, including the way points were allocated to wins (3), losses (-1), and ties (0). As in experiment 1, this point allocation does not impact participants' optimal strategy in the task. We compare human performance against each of the bot opponents to understand the complexity of patterns participants can modify in their *own* moves.

2.3.3 Bot Opponent Behavior

Participants were paired with one of eight adaptive bot opponents for the duration of the experiment. Each bot had an identical decision policy of choosing the move that would beat whatever move it estimated was most likely for its human opponent in the next round. In the event that multiple opponent moves were considered equally probable, the bot sampled one at random and chose the move that beat the sampled move.[5] What differentiated the bots across conditions was *how* they determined their human opponent's most likely move over consecutive rounds—each bot relied on a particular sequential dependency in its opponent's choices to predict their next move. To illustrate, a naïve

[4] The data and code for this experiment and for all analyses reported here can be found on github at: https://github.com/erik-brockbank/rps-bot-manuscript-public.

approach would involve simply tracking a participant's cumulative proportion of *rock*, *paper*, and *scissors*, then selecting the move each round that would beat whichever opponent choice had been most frequent so far. Given the simplicity of this dependency and the fact that people are unlikely to show a strong ongoing bias towards a particular move, we would not expect such a bot to perform particularly well against human opponents. However, tracking more complex patterns in participants' moves presents an opportunity for more successful prediction and exploitation.

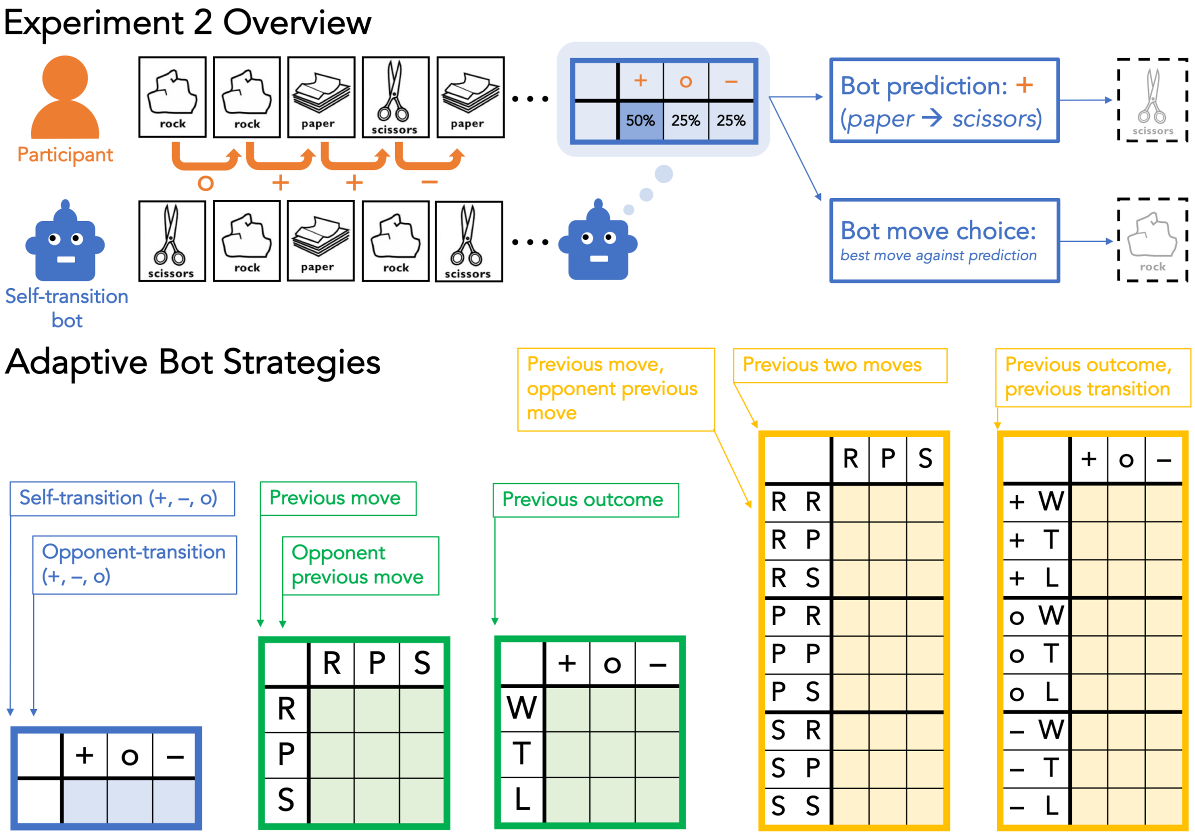


Figure 2.5. Experiment 2 overview.

The eight adaptive bots in the current experiment tested this idea by exploiting a broad range of behaviors in their human opponents. These were chosen to encompass the full scope of sequential regularities exhibited by the stable bots in experiment 1. This allows for direct comparison between the complexity of patterns in opponent behavior that people successfully adapted to and those they are able to minimize in their own moves. Further, these eight dependencies comprise various sequential regularities previously observed in games among human dyads (Brockbank & Vul, 2020), so all of them are good candidates for potentially exploiting participants over many rounds. For example, the first bot described below (`Self-transition`) tracks the proportion of moves its opponent makes corresponding to each *self-transition* (+, −, 0). If participants demonstrate any bias towards self-transitions (like those they successfully adapted to against the `Previous move (+)` and `Previous move (−)` bots in *experiment 1*), this bot will attempt to exploit such a bias. The bot's success against participants therefore provides an indication of how rigidly people exhibit any kind of stable self-transition bias in their moves. Each of the eight adaptive bot strategies are illustrated in Figure 2.5 and described in detail below in order of increasing complexity.

- **Self-transition** bot, described above, tracks participant *self-transitions* and chooses the move each round which beats its opponent's most likely self-transition (+, −, 0).
- **Opponent-transition** bot chooses a move based on the participant's most likely *opponent-transition* (+, −, 0), i.e., the participant's most likely transition relative to the bot's previous move.
- **Previous move** bot tracks the co-occurrence of every combination of participant moves from one round to the next. For example, does the participant tend to play *rock* after playing *paper*? This is similar to the `Self-transition` bot but allows for the possibility that a participant's most likely self-transition may vary at times depending on their previous move (i.e., if they choose *rock* most often after *rock*, but *paper* most often after *scissors*, these represent self-transition biases that differ across prior moves).

[5] Bots were indifferent between moves that had a 50/50 chance of a win or tie, and those that had a 50/50 chance of a win or loss. In this way, the bots maximized expected *win count*, but did not maximize expected *win count differential* (e.g., by favoring ties over losses).

- **Opponent previous move** bot is identical to the `Previous move` bot above, except that it exploits any pattern in participant moves based on their *bot opponent's* previous move rather than their own.
- **Previous outcome** bot tracks a participant's most likely transition conditioned on each previous outcome. "Win-stay, lose-shift" behavior or any other *outcome-transition* dependency will be exploited by this bot.
- **Previous move, opponent previous move** bot tracks player move choices given each *combination* of their own previous move *and their bot opponent's* previous move. This bot relies on the same information as the `Previous outcome` bot but once again encodes the dependency more richly by tracking all unique combinations of the two players' prior moves and the participant's subsequent move.
- **Previous two moves** bot chooses the move which beats its human opponent's most likely move given the participant's move choices in each of the previous two rounds.
- **Previous outcome, previous transition** bot exploits any dependency participants exhibit in their transitions each round, given both the outcome of the previous round and the transition they made in the previous round. For example, if participants were more likely to shift *up* after a round in which they shifted *up* and *won*, this bot will detect such a pattern and exploit it; this bot adapts to patterns like those exhibited by the *dual outcome-prior-transition* bot in experiment 1.

The adaptive bot strategies above capture a broad range of prior events that might impact a participant's move choices in predictable ways: their own previous move (`Self-transition` and `Previous move`), their opponent's previous move (`Opponent-transition` and `Opponent previous move`), the two combined (`Previous outcome` and `Previous move, opponent previous move`), the participant's previous *two* moves (`Previous two moves`), or the participant's previous two moves alongside the opponent's previous move

(`Previous outcome, previous transition`). By comparing participant behavior against each bot, we obtain a precise measure of people's ability to adaptively modify their own actions to avoid being exploited on any of these dimensions.

2.3.4 Adaptive Bot Complexity

Intuitively, the eight adaptive bots described above differ in the *complexity* of their strategies, with some tracking simple regularities in their opponent's moves and others relying on a rich set of contingencies across rounds. We can quantify the complexity of a bot's strategy based on the *memory demands* of tracking the dependency it uses to predict its opponent (see Figure 2.5). The simplest adaptive transition bots (`Self-transition` and `Opponent-transition`) need only store and update three counts in memory: a $1x3$ matrix with the number of $+$, $-$, and 0 transitions they observe. Meanwhile, the next three strategies above (`Previous move`, `Opponent previous move`, and `Previous outcome`) have an *intermediate* complexity because they instead maintain *nine* counts in memory using a $3x3$ matrix of transition or move frequencies based on additional information from the previous round (i.e., each possible previous move or previous outcome). Finally, the most complex bots (`Previous move, opponent previous move`, `Previous two moves`, and `Previous outcome, previous transition`) rely on a $9x3$ (*27 count*) matrix which tracks opponent moves or transitions given nine possible previous event combinations (two previous moves or a previous transition and previous outcome) to estimate their opponent's most likely move. While this is not the only way to characterize the complexity of these strategies, the memory required to generate opponent predictions offers a straightforward description of how the strategies vary, which could account for differences in people's success against the bots. We therefore explore how people's ability to adapt to each bot changes with the complexity of the dependency the bot exploits.

2.3.5 Results

Performance against adaptive bots

How successful was each of the adaptive bot strategies against human opponents? In experiment 1, we plotted participant win percentages (Figure 2.3) to illustrate how well participants exploited the stable patterns in their bot opponents' moves. Here, we instead examine average bot *win count differentials* across conditions (see Figure 2.6). The bot's win count differential is the number of times it beat its human opponent minus the number of times the human opponent won; this indicates how effectively each bot was able to exploit its human opponents, in essence revealing the degree to which participants were *trapped* in particular move patterns. Values greater than zero suggest that participants were reliably exploitable in that condition, while values close to zero indicate chance performance. Values less than zero result from participants successfully *counter-exploiting* their bot opponent. The bot win count differentials in Figure 2.6 illustrate a clear relationship between each bot's strategy complexity and how effectively that bot exploited participants.

First, the three bot strategies which tracked the most complex dependencies in human move choices (27-cell memory) were able to consistently beat participants. Bot win count differentials were significantly greater than zero for all of these strategies (`Previous move, opponent previous move`: $t(24) = 2.44$, $p = .02$; `Previous two moves`: $t(19) = 4.36$, $p<.001$; `Previous outcome, previous transition`: $t(25) = 5.68$, $p<.0001$). For the two bot strategies with the highest average win count differentials, only 4 out of 20 (`Previous two moves`) and 4 out of 26 (`Previous outcome, previous transition`) participants had win count differentials greater than or equal to zero; this is significantly fewer than would be expected by chance (binomial test $p = .01$ and $p<.001$, respectively) and suggests that most people paired with these bots would not be expected to come out ahead over many rounds. In short, participants exhibited the most complex regularities enough for the adaptive bots to exploit them, and further, people were essentially trapped in these behavior patterns, indicating a clear limit to the structure they can minimize in their own moves.

In the three *intermediate* (9-cell memory) bot conditions, participants were far less exploitable. Two of the bots obtained win count differentials which were not significantly different from zero (`Previous move`: $t(23)$ = -1.63, p = .12; `Previous outcome`: $t(21)$ = 0.07, p = .95) while bot win count differentials for the third were significantly *less* than zero (`Opponent previous move`: $t(27)$ = -3.11, p<.01). Intriguingly, given the evidence from prior work that people often exhibit "win-stay, lose-shift" behavior over repeated rock, paper, scissors rounds (Baek et al., 2013; Brockbank & Vul, 2020; Dyson et al., 2018; Dyson et al., 2016; Wang et al., 2014), the current results suggest that this pattern of outcome-dependent responding may be tempered when it is being actively exploited by an opponent. Participants' apparent success at evading exploitation in this way is also interesting in light of their minimal ability to adapt to these same patterns in an opponent's moves in experiment 1. More generally, these results present a clear contrast with those of the 27-cell strategies above; participants managed to avoid any kind of systematic exploitation based on the previous outcome or their own prior move, leading to chance performance over the 300 rounds in these conditions.

Finally, bot win count differentials in the simplest 3-cell memory conditions (along with the `Opponent previous move` condition above) suggest highly successful adaptive behavior by *participants*. Bot win count differentials were significantly *less* than zero for the transition bots at the far left in Figure 2.6 (`Self-transition`: $t(20)$ = -2.30, p = .03; `Opponent-transition`: $t(25)$ = -6.09, p<.01). At an individual level, bot win count differentials were negative against 17 out of 21 (`Self-transition`) and 22 out of 26 (`Opponent-transition`) participants in these conditions, significantly more than would be expected by chance (binomial test, p<.01 and p<.001, respectively). These results suggest that participants reliably *outwitted* bot strategies that exploit simple transition regularities. The only way to achieve this pattern of results is for participants to have discovered a way of counter-exploiting the bot opponents. *How did they accomplish this?*

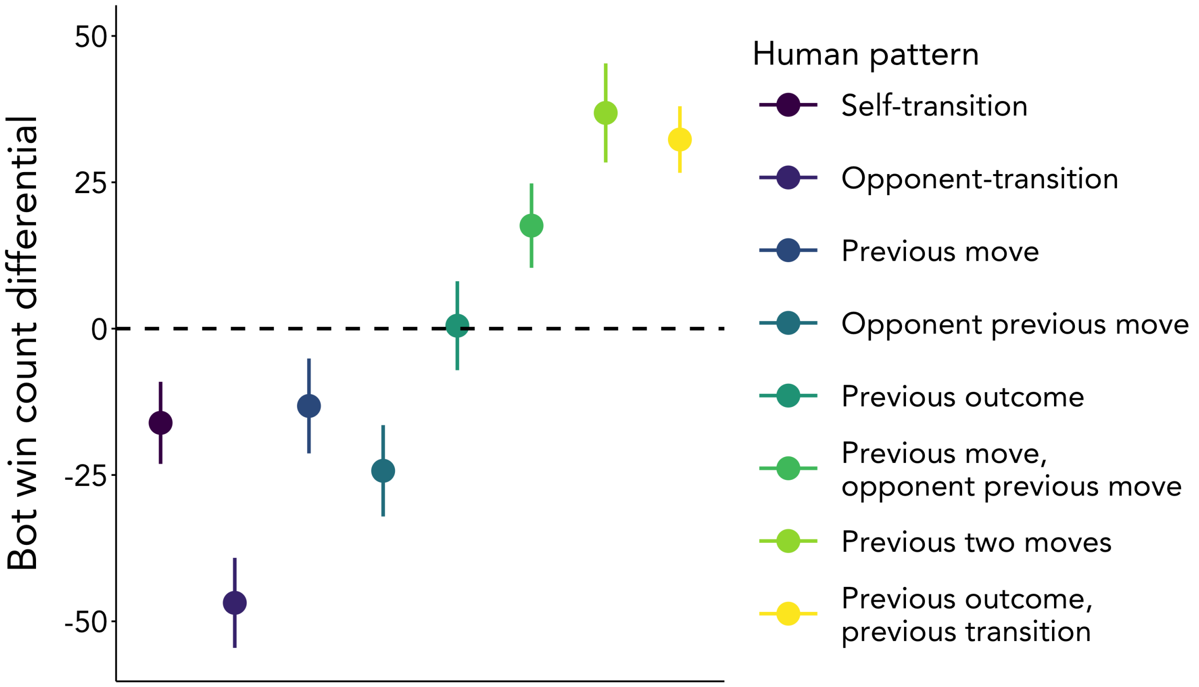


Figure 2.6. Success of adaptive bot strategies against human opponents. Dashed line indicates chance performance; error bars show standard error of the mean. Bot win count differentials greater than chance reveal strategies that reliably outsmart human opponents, while values less than chance indicate successful counter-exploitation by human players. Participants were unable to eliminate complex regularities in their own behavior (far right), but effectively counter-exploited the simpler bot strategies (far left).

Strategic responding to adaptive bots

The bot win count differentials in Figure 2.6 illustrate a surprising result: Bots that chose their moves based on participants' most likely self-transitions, opponent-transitions, or, relatedly, based on participant move choices given their opponent's prior move, *lost systematically*. In short, people in these conditions selected actions that successfully *counter-exploited* their bot opponents. What sort of strategic responding allowed them to outwit these adaptive bots? Here, we present exploratory analyses aimed at better understanding this question, focusing on the simplest (3-cell memory) transition bots.

Given participants' successful adaptation to stable *transition* patterns in experiment 1, we investigate whether they might have leveraged similar transition biases in their *own* moves to counter-exploit the adaptive bots here. Specifically, we examine evidence for strategic use of *self-transition* biases against the *opponent-transition* tracking bot, and *opponent-transition* biases against the *self-transition* tracking bot. In each case, we rely on simulated game play to understand whether a particular transition bias had a *theoretical* advantage against the relevant bot opponent; we then evaluate empirical behavior to understand the extent to which people did in fact capitalize on the theoretical advantages of a given transition bias. Our analyses suggest that there was a clear theoretical advantage to favoring particular *self-transitions* against the *opponent-transition* adaptive bot and that participants' highly successful counter-exploitation of this bot reflected strategic use of self-transition biases. However, the source of participant's more moderate success against the self-transition-exploiting adaptive bot remains less clear; here, *opponent-transition* biases were not theoretically adaptive. Nonetheless, we do find evidence that participants favored particular opponent-transitions against this bot, suggesting a potentially adaptive use of these biases that our simulations are unable to detect.

In experiment 1, participants reliably beat the stable self- and opponent-transition bots (Figure 2.3); in these conditions, a player's best move was always a particular transition relative to their own previous move. Thus, the previous experiment provides evidence that people flexibly incorporated self-transition biases into their own moves when doing so was adaptive. Might a similar strategy have worked in the current experiment? We first examine whether exhibiting a strong *self-transition* bias could have performed well against the bot that exploited *opponent-transition* dependencies, since participants showed the greatest success against this bot (and a stable self-transition bias would *not* have been effective against the `Self-transition` bot). To do this, we ran *simulated matches* between the bot and 1000 simulated human players that exclusively chose a single self-transition for 300 rounds.[6] We ran separate 1000-person, 300-round simulations for each possible self-transition bias (+, −, or 0) against the `Opponent-transition` bot. We then explored the outcomes for the 1000 "participants" in each simulation. Did favoring a particular self-transition ever lead to systematic counter-exploitation of the bot that tracked opponent-transition patterns?

Against the bot that exploited participants' *opponent-transitions*, there was a distinct potential advantage to participants for favoring a single *self-transition*. A stable self-transition bias yielded a positive win count differential of 0.2 per trial for approximately 36% of simulated participants (across the 300 rounds, this corresponded to a total win count differential between 40 and 80). For most of the other simulated participants, it led to an average win count differential of 0, with a small fraction (roughly 10%) having a win count differential of -1 per trial (i.e., losing every game). This latter outcome is dramatic, but likely reflects the fact that for any two starting rounds in which the bot beats the simulated participant both times (1/9 probability), subsequent reliance on a particular *self-transition* by the simulated participant reinforces a stable *opponent-transition* pattern that the bot can continually exploit. These results were similar for all three self-transition biases (*up*, *down*, or *stay*). In other words, while favoring a given self-transition was not *always* adaptive against the `Opponent-transition` bot, doing so could sometimes maneuver the bot into move sequences that led to a substantially greater number of wins and ties than losses for the simulated participant. This suggests one possible mechanism by which our actual participants may have effectively counter-exploited the `Opponent-transition` bot (Figure 2.6). Critically, because it is not *always* advantageous, it requires that participants implement such a counter-exploitation strategy *adaptively*, i.e., only when doing so produces consistently higher win rates.

Could favoring a particular transition lead to similar success against the `Self-transition` bot? Exhibiting a stable *self-transition* bias would evidently not be effective,

[6] Code and results for all simulations can be found on github at: https://github.com/erik-brockbank/rps-bot-manuscript-public.

but it may be that an *opponent-transition* bias such as copying the bot's previous move or playing the move that would beat its previous move was helpful. We ran an identical simulation as above with 1000 simulated participants playing 300 rounds against the `Self-transition` bot. This time, simulated participants always chose a particular *opponent-transition* (+, −, or 0). Here, we find that favoring a single opponent-transition did *not* produce conditions for successfully counter-exploiting the bot. For roughly 40% of simulated subjects, choosing a stable opponent-transition produced a per-trial win count differential very close to 0 (all of these were positive but never more than 0.01). Most of the remainder had a win count differential around -0.5 per trial, with roughly 10% obtaining an average win count differential of -1 (as before, this reflects participant and bot choices across the first two rounds wherein participants subsequently favoring an opponent-transition also maintains a stable self-transition dependency the bot can exploit). As with the self-transition simulations, this pattern held for all opponent-transition biases (*up*, *down*, or *stay*). Thus, while favoring a particular *opponent-transition* against the `Self-transition` bot might have allowed some participants to avoid exploitation by this opponent, it is not clear that such a strategy could have enabled participants to successfully *counter-exploit* it as they did (Figure 2.6).

In sum, there was a theoretical advantage for players adopting *self-transition* biases against the `Opponent-transition` bot; against the `Self-transition` bot, any counter-exploitation achieved by generating *opponent-transition* biases would have required a more nuanced strategy than our simulations captured. However, it remains an open question whether modifying self- or opponent-transition biases is *in fact what participants did*. To test the degree to which participants relied on stable transition patterns to counter-exploit their opponents, we evaluate whether they showed larger *self-transition* biases against the bot that exploited *opponent-transitions* and increased *opponent-transition* biases against the bot that exploited *self-transitions*.

The extent to which participants demonstrated a particular dependency over 300 rounds can be precisely quantified in terms of the *information gain* for that dependency in their moves (Brockbank & Vul, 2021). Briefly, information gain measures how much the distribution of moves for a given dependency deviates from uniform. We use this measure to compare how much participants expressed self-transition and opponent-transition biases against different adaptive bot opponents. For example, given a sequence of participant move choices X, the probability distribution over self-transitions $T = \{+, -, 0\}$ will be the number of times each self-transition appeared in X divided by $|X|$. Intuitively, the more uneven this distribution is, the more participants exhibited an exploitable self-transition bias in X. The *Shannon entropy* H (Shannon, 1948) of the distribution formalizes this intuition: The *lower* the H value (less entropy) for a particular dependency, the *more* evident that dependency is in the underlying move sequence. The information gain $IG(T)$ is simply the distance between the Shannon entropy of a uniform distribution over transitions and the entropy of the empirical distribution, $H(T)$:

$$IG(T) = -\log(1/3) - H(T) = \log(3) + \sum_{i \in T} p(T_i) \ \log(p(T_i))$$

This value will be larger the more non-uniform the distribution of transitions T is; in short, IG quantifies how much a participant's moves were predictable via a given dependency, with larger values indicating that people demonstrated that regularity more.

How much did participants display stable *self-transition* and *opponent-transition* biases against each of the adaptive bots? First, as above, we consider the strategic use of *self-transitions.* Figure 2.7 (Top) shows the information gain for self-transition dependencies exhibited by participants against each of the different adaptive bot opponents. Here, a clear pattern stands out. Overall, participants showed very little self-transition regularity, particularly against bot opponents that would have exploited any such regularity (`Self-transition` and `Previous choice`). However, participants exhibited the highest self-transition information gain in their moves against bots that tracked *opponent-transition*

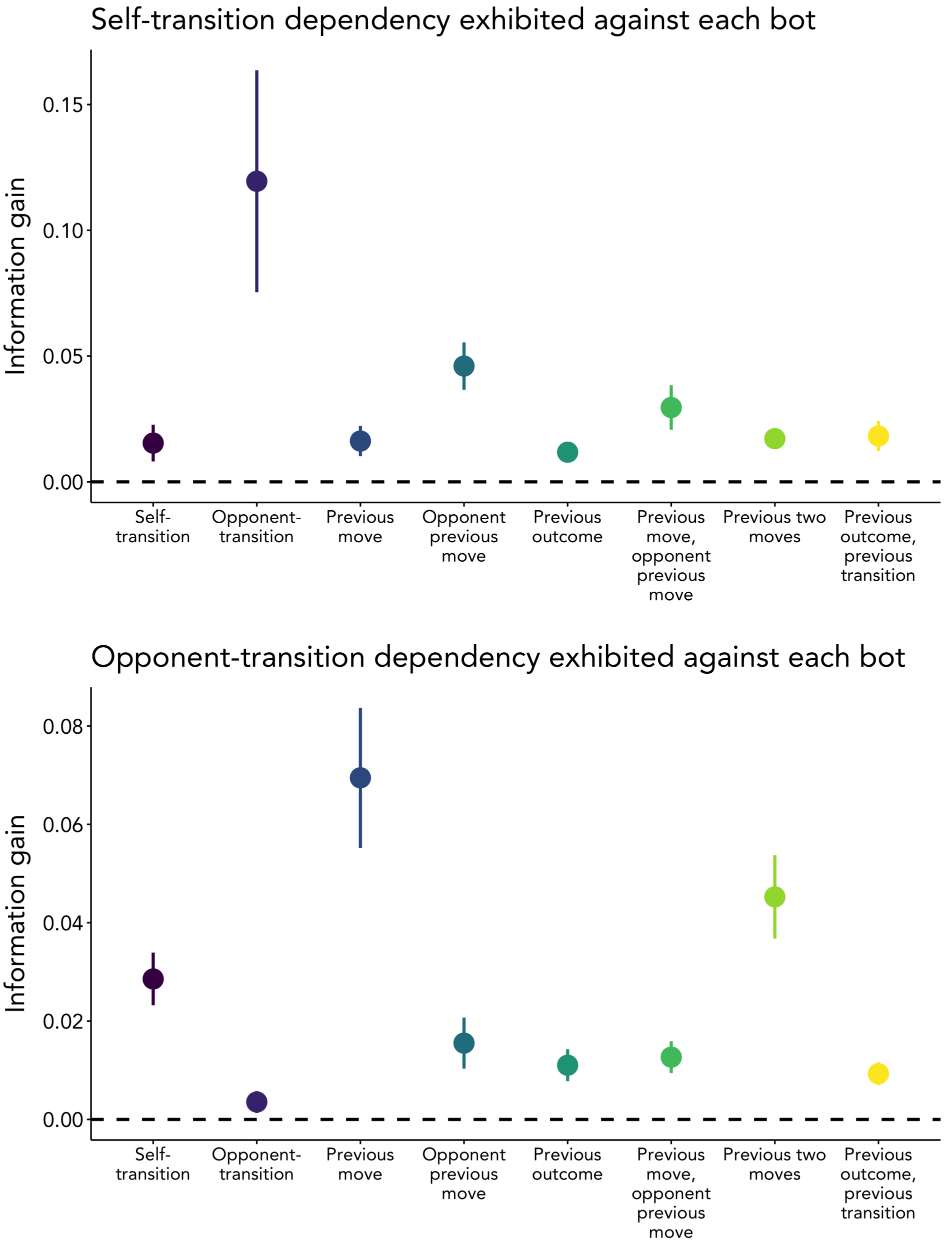


Figure 2.7. Participants' exploitability for self- and opponent-transitions against each bot opponent. (Top) Information gain for self-transition dependencies in participant move choices against each bot opponent. (Bottom) Information gain for opponent-transition dependencies in participant move choices against each bot. Dashed lines show chance performance and error bars reflect standard error of the mean. Participants showed evidence of modulating their self-transition and opponent-transition dependencies to reduce exploitability and counter-exploit adaptive bot opponents.

dependencies: `Opponent-transition` and `Opponent previous move` (and, to some degree, `Previous move, opponent previous move`). There is a significant difference in information gain for participants' self-transition dependency between the `Self-transition` and `Opponent-transition` bots ($t(45)$ = -2.10, p = .04), which are otherwise matched in the complexity of their strategies, as well as between the `Previous move` and `Opponent previous move` bots ($t(50)$ = -2.57, p = .01) which are similarly matched. Thus, participants paired with opponent-transition-exploiting bots appeared to rely on increased self-transition biases as a means to counter-exploit them. Consistent with our simulation results, this strategy appears to have been successful; participants had the highest average win count differential in the `Opponent-transition` and `Opponent previous move` conditions (Figure 2.6) where they also exhibited the greatest self-transition dependencies.

Do participants exhibit a similar effect for opponent-transition dependencies, namely a greater *opponent-transition* bias against the bots that tried to exploit *self-transition* patterns? Figure 2.7 (Bottom) shows information gain for participants' opponent-transitions against each of the adaptive bots; the pattern is qualitatively similar to Figure 2.7 (Top). Participants displayed little opponent-transition bias against the bots that exploited this pattern (`Opponent-transition` and `Opponent previous move`). In contrast, they exhibited greater opponent-transition dependencies against bots that tried to exploit *self-transition* biases (`Self-transition`, `Previous move`, and `Previous two moves`). In line with similar effects above, there is a significant difference in information gain for the opponent-transition dependency between the symmetrical `Self-transition` and `Opponent-transition` bots ($t(45)$ = 4.63, p<.001), as well as between the `Previous move` and `Opponent previous move` bots ($t(50)$ = -3.77, p<.001). Broadly, participants showed increased *opponent-transition* dependencies against bots that sought to exploit their *self-transition* biases, mirroring the pattern of increased self-transitions described above. Despite the similar overall pattern, the magnitude of participants' elevated opponent-transition biases (Figure 2.7 (Bottom)) is reduced relative to the self-transition biases

(Figure 2.7 (Top)). One reason for this may be that it was less adaptive. The simulation results described previously suggest that stable opponent-transition biases did not permit systematic counter-exploitation of an opponent tracking self-transition dependencies. Consistent with this, participants did not counter-exploit the self-transition tracking bots as effectively as participants paired with the opponent-transition tracking bots (Figure 2.6).

2.3.6 Discussion

In the current experiment, we tested people's ability to minimize sequential patterns in their move choices when being exploited by an adaptive opponent. Participants played 300 rounds of rock, paper, scissors against a bot that tried to predict the participant's most likely move based on prior moves and outcomes. We tested eight bot opponents that varied in the complexity of the patterns they relied on to predict their human adversary; these encapsulate the full range of stable dependencies exhibited by the bots in experiment 1, allowing for a direct comparison between the patterns people can adapt to in an opponent, and those they can revise in their own moves. We first assessed participants' overall success against the adaptive bots to understand how flexibly participants reduced exploitable dependencies in their own decisions. Next, we looked at the degree to which participants exhibited *self-transition* and *opponent-transition* dependencies against each bot to understand the basis for their adaptive behavior.

Our results contain two central findings. First, overall performance against the adaptive bots was consistent with the complexity of the behavioral pattern each bot exploited. Against the most sophisticated adaptive bots, participants lost reliably over the course of the 300 rounds. However, paired with bots that merely tried to exploit simple transition dependencies in their opponent's actions, participants showed evidence of successfully *outwitting* them. Our second key finding concerns the nature of this counter-exploitation. We find that participants exhibited increased *self-transition* dependencies when paired with bots that exploited *opponent-transition* biases, and vice versa with bots that exploited self-transition biases. Though these results were exploratory, they suggest that participants' success against the simpler adaptive bots primarily revolved around their ability to modify self-transition and opponent-transition patterns in their own moves. However, our simulation results provide important context for these findings. Against the bot that exploited opponent-transition biases, favoring a particular self-transition led to systematic advantages for a subset of simulated participants; thus, participants' empirical success against this bot must have relied on *adaptive* use of self-transition biases, employing them only when advantageous. Against the self-transition tracking bot, participants' successful counter-exploitation, though less dramatic (Figure 2.6), is also harder to explain, since our simulation results did not find a stable advantage for the opponent-transition biases seen in participants' empirical behavior.

Broadly, these results suggest that people are flexible in their use of *self-transition* and *opponent-transition* patterns to adapt to a strategic opponent. And their use of these patterns is unlikely to be random or thoughtless, since doing so is not always advantageous. However, this adaptive ability appears limited to a choice among these relatively simple transition-level behavioral dependencies. Against bots that exploited more complex patterns, participants were essentially "stuck in their ways" and lost reliably. Changes to one's own behavior in this adversarial setting enabled people to effectively counter-exploit simple opponents, but lacked the scope needed to adapt to more complex opponents.

2.4 General Discussion

In this work, we address the question of how people perform *adaptive* reasoning in an adversarial setting. Across two experiments using the game of rock, paper, scissors, participants demonstrated a highly selective ability to exploit patterns in their opponent's moves and revise stable dependencies in their own moves. These adaptive responses exhibited a consistent reliance on detecting and modifying *transitions* from one move to the next, but little ability to generalize to more complex patterns. In this way, results paint a clear picture of people's sequential, adaptive reasoning as flexibly utilizing a well-defined but surprisingly limited set of behaviors.

In experiment 1, participants played 300 rounds of rock, paper, scissors against one of seven bot opponents, each of which chose its moves according to a different sequential dependency. Bot opponents varied in the complexity of their *strategies* based on the number of prior events that determined their moves: the simplest *self-transition* and *opponent-transition* bots followed reliable patterns in their transitions from one move to the next, while intermediate *outcome-transition* bots favored particular transitions *depending on the prior outcome* (e.g., "win-stay, lose-shift"), and the most complex *dual outcome-prior-transition* bot chose a different transition depending on each unique combination of prior outcome and prior transition. We measure participants' ability to successfully exploit each of these stable patterns in their opponent's moves to better understand the scope of sequential dependencies that people can adapt to in this setting. We find that participants are highly successful against *transition-level* patterns over 300 rounds but struggle to adapt to more complex opponent behaviors. Despite a large body of work demonstrating people's tendency to exhibit *win-stay, lose-shift* responding in this and other settings (Brockbank & Vul, 2020; Dyson et al., 2018; Dyson et al., 2016; Wang et al., 2014), we find that the ability to adapt to this pattern in an opponent is largely limited to exploiting individual transitions (those following a particular outcome) rather than the full dependency structure. Broadly, the results suggest that the space of behavioral regularities people can reliably exploit in an opponent over many rounds is restricted to contingencies on either player's previous move. While such patterns can be readily exploited, additional events such as the previous outcome are mostly ignored, even when they are predictive.

In experiment 2, we examine whether the limitations in adaptive behavior that participants exhibited in experiment 1 extend to revising patterns in one's *own* actions. This addresses the possibility that people have an easier time monitoring sequential regularities in their own moves than exploiting them in others. Participants once again played 300 rounds of rock, paper, scissors against a bot opponent. However, this time the bot chose its moves in an effort to exploit different patterns that the *participants themselves* demonstrated. These patterns were chosen to align with the dependencies that the stable bots in experiment 1 displayed in their moves. We measure how reliably each bot won against its human opponents to understand which exploitable patterns people can eliminate in their own choices and which ones are outside their control. Our findings are consistent with those obtained in experiment 1. Against bots that exploited *self-transition* and *opponent-transition* patterns, participants won reliably over the 300 rounds. Follow-up analyses suggest that they did so by modifying self-transition and, to a lesser degree, opponent-transition patterns in their own moves to outwit the bots. However, their adaptive behavior was limited to this level of strategic complexity; participants lost reliably to bots that exploited more complex sequential patterns. Broadly, results from experiment 2 suggest that people can flexibly and adaptively modify the ways in which they transition from one move to the next or respond to an opponent's prior move, but show little ability to revise their own actions in more sophisticated ways.

Results from experiment 2 reinforce and complement the findings from experiment 1. In experiment 1, participants were highly successful against stable *transition-level* patterns in their opponent's moves; successfully exploiting these bots required that participants implement self-transition biases in their own moves. We also find in experiment 1 that participants' success against the *outcome-transition* bots is primarily restricted to adaptive responding after a *tie* outcome. This pattern of results can further be explained by participants exhibiting stable self-transition biases, which leads to fewer losses for the participants and disproportionate wins after a tie against the *outcome-transition* bots. However, the increased self-transition biases in experiment 1 might have resulted primarily from people recognizing predictable patterns in their opponent's moves and responding accordingly. Results from experiment 2 show that people also manipulate transition biases in their moves as a means of counter-exploiting an adaptive opponent. Collectively, findings suggest that in repeated rock, paper, scissors interactions, people will successfully outwit their opponent when doing so involves detecting or modifying simple move patterns. However, this flexibility is restricted to a constrained set of such patterns. In essence, the *hypothesis space* of structured behaviors people represent in this task appears to be quite small.

Beyond adversarial interactions, the present results are informative for broader questions about the nature of human social learning. The rock, paper, scissors game, with its simple rules and lack of prior knowledge or strategy, represents a simplified and distilled setting for reasoning about sequential structure in another person's behavior. *Why then, did participants struggle to adapt to all but the most basic patterns in their opponent's moves and their own?* One possible answer is that people's ability to reason about sequential regularities in others' actions is highly constrained and the current results simply expose those underlying limitations. However, a growing body of work suggests that in many social learning settings, people rely on rich mental models of others to predict and understand their behavior in nuanced ways (Vélez & Gweon, 2021). The capacity for this learning relies on having a sophisticated *theory of mind* that forms the basis of complex inferences about why others act the way they do (Baker et al., 2009; Jara-Ettinger et al., 2016). This work offers a potentially useful lens for interpreting the current results.

Critically, traditional theory of mind reasoning is difficult to instantiate in the rock, paper, scissors game for several reasons. First, the adversarial nature of the interaction injects ambiguity into the possible causes of an opponent's behavior—are patterns in their moves a result of naïveté or a trap by a more sophisticated reasoner? More importantly, the structure of the game makes it difficult to predict an opponent's moves on the basis of inferences about their mental states; instead, a player's primary recourse is to search for exploitable patterns in their opponent's moves across repeated rounds. The current results suggest that people's adversarial reasoning in a setting with such impoverished mental-state attributions is highly limited. Understanding the source of these limitations presents a clear opportunity for future work. Inferences about others' behavior may not always rely on a complete theory of mind (Burger & Jara-Ettinger, 2020), such as when we infer that someone is acting out of habit (Gershman et al., 2016). When features of the task or the environment limit the availability of robust theory of mind reasoning, how do people interpret and adapt to others' behavior? The current results offer one means by which to address this question.

In sum, this work explores how people adapt in repeated adversarial interactions. What kind of structured behavioral patterns can people predict and exploit in others? And how well can they modify these patterns in their own actions to avoid similar exploitation? By observing decision-making across repeated rounds of mixed strategy equilibrium gameplay, we obtain a clear and precise account of the limits of people's adaptive adversarial behavior.

2.5 Acknowledgments

I thank my co-author Edward Vul for his hallowed and unassailable intellect tirelessly applied to the many earlier drafts, analyses, and revisions of this manuscript.

Chapter 2, in full, has been submitted for publication of the material and is currently in revision as it may appear in *Cognitive Psychology.* An earlier version of the project was published as Brockbank, E., & Vul, E. (2021). Humans fail to outwit adaptive rock, paper, scissors opponents. In T. Fitch, C. Lamm, H. Leder, & K. Teßmar-Raible (Eds.), *Proceedings of the 43rd Annual Meeting of the Cognitive Science Society* (pp. 1740-1746). Cognitive Science Society. The dissertation author was the primary investigator and author of this material.

References

Axelrod, R. (1984). *The evolution of cooperation.* Basic Books.

Baek, K., Kim, Y. T., Kim, M., Choi, Y., Lee, M., Lee, K., Hahn, S., & Jeong, J. (2013). Response randomization of one-and two-person rock-paper-scissors games in individuals with schizophrenia. *Psychiatry research*, *207*(3), 158–163.

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition*, *113*(3), 329–349.

Bar-Hillel, M., & Wagenaar, W. A. (1991). The perception of randomness. *Advances in applied mathematics*, *12*(4), 428–454.

Bonawitz, E., Denison, S., Gopnik, A., & Griffiths, T. L. (2014). Win-stay, lose-sample: A simple sequential algorithm for approximating bayesian inference. *Cognitive psychology*, *74*, 35–65.

Brockbank, E., & Vul, E. (2020). Recursive adversarial reasoning in the rock, paper, scissors game. In S. Denison, M. L. Mack, Y. Xu, & B. C. Armstrong (Eds.), *Proceedings of the 42nd annual conference of the cognitive science society* (pp. 1015–1021). Cognitive Science Society.

Brockbank, E., & Vul, E. (2021). Formalizing opponent modeling with the rock, paper, scissors game. *Games*, *12*(3), 70.

Budescu, D. V., & Rapoport, A. (1994). Subjective randomization in one- and two-person games. *Journal of Behavioral Decision Making*, *7*(4), 261–278.

Burger, L., & Jara-Ettinger, J. (2020). Mental inference: Mind perception as bayesian model selection. *CogSci*.

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

Camerer, C., & Weigelt, K. (1988). Experimental tests of a sequential equilibrium reputation model. *Econometrica: Journal of the Econometric Society*, 1–36.

Dyson, B. J. (2019). Behavioural isomorphism, cognitive economy and recursive thought in non-transitive game strategy. *Games*, *10*(3), 32.

Dyson, B. J., Steward, B. A., Meneghetti, T., & Forder, L. (2020). Behavioural and neural limits in competitive decision making: The roles of outcome, opponency and observation. *Biological psychology*, *149*, 107778.

Dyson, B. J., Sundvall, J., Forder, L., & Douglas, S. (2018). Failure generates impulsivity only when outcomes cannot be controlled. *Journal of Experimental Psychology: Human Perception and Performance*, *44*(10), 1483.

Dyson, B. J., Wilbiks, J. M. P., Sandhu, R., Papanicolaou, G., & Lintag, J. (2016). Negative outcomes evoke cyclic irrational decisions in rock, paper, scissors. *Scientific Reports (Nature Publisher Group)*, *6*(1), 20479.

Forder, L., & Dyson, B. J. (2016). Behavioural and neural modulation of win-stay but not lose-shift strategies as a function of outcome value in rock, paper, scissors. *Scientific Reports (Nature Publisher Group)*, *6*(1), 33809.

Gershman, S. J., Gerstenberg, T., Baker, C. L., & Cushman, F. A. (2016). Plans, habits, and theory of mind. *PloS one*, *11*(9), e0162246.

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological review*, *103*(4), 650.

Griffiths, T. L., Daniels, D., Austerweil, J. L., & Tenenbaum, J. B. (2018). Subjective randomness as statistical inference. *Cognitive psychology*, *103*, 85–109.

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.

Kangas, B. D., Berry, M. S., Cassidy, R. N., Dallery, J., Vaidya, M., & Hackenberg, T. D. (2009). Concurrent performance in a three-alternative choice situation: Response allocation in a rock/paper/scissors game. *Behavioural Processes*, *82*(2), 164–172.

Lie, C., Baxter, J., & Alsop, B. (2013). The effect of opponent type on human performance in a three-alternative choice task. *Behavioural Processes*, *99*, 87–94.

Lopes, L. L., & Oden, G. C. (1987). Distinguishing between random and nonrandom events. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *13*(3), 392.

Moulin, H. (1986). *Game theory for the social sciences*. NYU press.

Nash, J. F. (1950). Equilibrium points in n-person games. *Proceedings of the national academy of sciences*, *36*(1), 48–49.

Palacios-Huerta, I. (2003). Professionals play minimax. *The Review of Economic Studies*, *70*(2), 395–415.

Rapoport, A., & Budescu, D. V. (1992). Generation of random series in two-person strictly competitive games. *Journal of Experimental Psychology: General*, *121*(3), 352.

Rapoport, A., & Budescu, D. V. (1997). Randomization in individual choice behavior. *Psychological Review*, *104*(3), 603–617.

Rapoport, A., & Chammah, A. M. (1970). *Prisoner's dilemma: A study in conflict and cooperation* (Vol. 165). University of Michigan press.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell system technical journal*, *27*(3), 379–423.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al. (2016). Mastering the game of go with deep neural networks and tree search. *nature*, *529*(7587), 484–489.

Stöttinger, E., Filipowicz, A., Danckert, J., & Anderson, B. (2014). The effects of prior learned strategies on updating an opponent's strategy in the rock, paper, scissors game. *Cognitive Science*, *38*(7), 1482–1492.

Tversky, A., & Kahneman, D. (1972). Subjective probability: A judgment of representativeness. *Cognitive psychology*, *3*(3), 430–454.

Vélez, N., & Gweon, H. (2021). Learning from other minds: An optimistic critique of reinforcement learning models of social learning. *Current Opinion in Behavioral Sciences*, *38*, 110–115.

Walker, M., & Wooders, J. (2001). Minimax play at wimbledon. *American Economic Review*, *91*(5), 1521–1538.

Wang, Z., Xu, B., & Zhou, H. J. (2014). Social cycling and conditional responses in the rock-paper-scissors game. *Scientific reports*, *4*, 5830.

Chapter 3

People update their mental models of other agents' knowledge to support collaborative physical judgments

INTERIM SUMMARY

In the previous two chapters, we first argued that repeated interactions in mixed strategy equilibrium games like rock, paper, scissors offer a unique venue for studying people's *behaviorist intuitive psychology.* We showed evidence that people build up predictive models of their opponents over many rounds of play which allow some people to exploit their opponent reliably. In the previous chapter, we took a systematic approach to understanding how people might go about doing this. Across two studies, participants were paired with algorithmic opponents that exhibited a range of stable patterns in their moves, or sought to exploit the same structures in participant moves. We analyze participants' success against these bots to understand what kind of opponent models people can develop in this setting and what kinds of patterns they can revise or limit in their own actions. Intriguingly, results suggest that people show a highly constrained capacity for adapting to complex patterns in their opponents. Our second experiment rules out the possibility that this is due to a mismatch between patterns they can exploit in others and patterns they can monitor in their own moves. Broadly, people seem unable to deploy highly sophisticated statistical learning in service of choosing actions in this setting.

Is this because people are simply bad at recognizing structure in other people's behavior more generally? This seems unlikely. Much of the research on *cognitivist intuitive psychology* described at the outset suggests that in fact, people can build remarkably rich and complex representations of others based on their behavior. In the next chapter, we investigate this process while taking a methodologically similar approach to chapters 1 and 2. Over a series of repeated *collaborative* interactions with a stable bot partner, participants receive suggestions in a simple physics-based game. Here, instead of varying the sequential patterns in the bot's behavior, we manipulate latent parameters that affect the bot's suggestions in recognizable ways. In experiment 1, we modify the *variance* of the bot's suggestions. In experiment 2, we manipulate the bot's own internal model of the task in a way that produces systematic *biases* in the bot's suggestions. Thus, while this work is framed quite differently from that in chapters 1 and 2, we retain the same conceptual approach of investigating repeated interactions with a partner whose decisions are parametrically manipulated in ways that cover a large space of structured behavior. Just as we used people's task success to probe their representations of these behaviors in the previous chapters, here we use their *interventions* on their bot partner's suggestions. In the next chapter, I argue that this approach, applied to people's collaborative behavior with bots that vary in their *competence*, allows once again for a precise characterization of people's *underlying representations of their bot partner* on the basis of previous interactions.

Abstract

How do people estimate the abilities of others across repeated interactions? We explore this question in a collaborative physical reasoning task with artificial agents whose underlying competence changes over time and whose knowledge reflects different internal models of the task itself. This approach builds on prior work by exploring people's *dynamic* social inferences along with their ability to represent the *internal model* of another agent in a physical task domain. Across two studies, participants played a physics-based video game paired with an agent who suggested moves on every round. We measured participants' decisions to accept or revise their partner's suggestions to understand how people appraised their partner's ability from one round to the next. In study 1, agent partners varied in whether they were *reliable*, *unreliable*, *improving*, or *worsening*. We find that participants were sensitive to these changes in ability and that their intervention behavior reflected latent estimates of their partner's competence above and beyond the accuracy of their partner's suggestion on a given round. In study 2, we build on these results by probing people's ability to learn a stable feature of their partner's physical world model, rather than a simple latent parameter like accuracy. Results suggest that people successfully inferred this underlying cause of their partner's errors and could systematically correct for it. Together, the findings offer a precise account of how people assess and update their estimate of another agent's abilities over time and integrate this information into their own behavior in collaborative settings.

Keywords: social learning, artificial agents, theory of mind, competence, trust

3.1 Introduction

Imagine being at a park with your young child and they ask if they can play on the monkey bars by themselves. How do parents estimate whether this is something their child can do safely? They might consider first and foremost whether their child has played on monkey bars in the past; or, they might search for other activities their child has done which require similar levels of coordination and may transfer to this scenario. They might refer to conversations they've had with other parents (or, if the child has an older sibling, whether the sibling was using monkey bars at this age). Or, they may reflect for a moment on whether the child's appraisal of their *own* ability can be trusted in this setting, or ask the child questions about their plans to assess the child's beliefs about the nature of the task and whether these beliefs seem accurate. Thus, even a simple evaluation about what might be safe behavior on the playground can, in principle, recruit a wide range of cognitive strategies and mental inferences. In everyday life, reasoning about what others *know* or *are capable of* is commonplace. Adults do this when coordinating on physical tasks like moving furniture or when discussing whether somebody else is a good party planner or dog sitter; teachers do this with their students; and children do this with each other and with adults around them to determine who they can trust. *How do we represent the competence of others?* The current work explores this question using repeated interactions in a collaborative, physics-based task, allowing for precise estimates of how people represent the abilities and underlying mental models of those around them.

Work on evaluations of competence sits at the intersection of several exciting areas of research spanning social psychology, robotics and artificial intelligence, and developmental psychology. Perhaps the most immediately relevant thread has explored people's *advice-taking* behavior; under what circumstances do people take others' advice and how do they incorporate what they know about their collaborator or the task into their subsequent actions? Work in this space has shown that people can integrate not just the content of others' advice but their confidence into joint decision making (Bahrami et al., 2010; Pescetelli & Yeung, 2021) similar to "cue integration" in sensory domains (Ernst & Banks, 2002). However, unlike integration of low-level sensory information, people's integration of others' advice is subject to systematic biases that can impact subsequent task performance, such as inflating others' expertise (Leong & Zaki, 2018). In addition, people appear to be differentially sensitive to the timing and valence of advice—prior work has shown "primacy" effects where early good advice is weighted more heavily (Staudinger & Büchel, 2013) and a tendency to weight good and bad advice differently (Biele et al., 2009). Despite the potential for such biases, the information that can be extracted from advice can be arbitrarily rich; people can in principle communicate not only their recommended response or their confidence but additional reasons for proferring the advice (e.g., their thought processes) and listeners can further incorporate rich inferences about their collaborator (e.g., their motivations or beliefs) into their behavior. However, work in this space has only recently begun to explore such inferences as part of the advice-taking process. For example, Vélez and Gweon (2019) find that when deciding whether to accept an algorithmic collaborator's advice in a simple card game, people's behavior reflects inferences not only about the advisor's overall helpfulness but also what the advisor knows and whether their recommendations tend to be *risky* or *conservative*.

A second prominent area where assessments of others' competence has risen to the fore is in evaluations of *trust*. Deciding who we can trust, and what we can trust them for, is a critical social evaluation for children and adults alike and, increasingly, must also be performed with respect to artificial agents. Interpersonal trust is a multi-faceted evaluation but relies at least in part on judgments of whether others are *capable* of helping in the first place. Prior work has shown that 4-5 year-olds' judgments of adults' competence are sensitive to both task-based cues such as difficulty as well as agent-based cues such as time to completion, but they struggle to integrate the two (Leonard et al., 2019); similarly, even toddlers differentiate other agents based on the time and effort required to complete a goal

(Jara-Ettinger et al., 2015). Together, this work suggests that even very young children have a relatively abstract (though still incomplete) representation of others' *competence* as a basis for who is trustworthy.

This ability to assess what others are capable of develops into adulthood, where a large body of work has explored the basis of people's trust in *artificial and robotic agents* (Chen et al., 2020; Soh et al., 2020; Xie et al., 2019). These investigations are largely motivated by a desire for artificial agents themselves to have an accurate internal model of when humans are likely to trust them. Broadly, this work emphasizes adults' ability to combine sophisticated judgments about the task and the agent when inferring their competence. For example, people appear to rely on latent task representations to infer that agents who are competent in one task are likely to be competent in similar tasks (Soh et al., 2020). Further, their judgments about agent abilities and collaborative decisions with these agents incorporate both accuracy and risk preference (Xie et al., 2019), as well as riskiness of the task itself relative to the agent's demonstrated abilities (Chen et al., 2020). Thus, like the monkey bars example at the outset, adults' decisions to collaborate with artificial agents combine past performance, ability on similar tasks, and latent traits like risk-aversion. However, as in the advice-taking literature, the information *conveyed* through robot performance tends to be sparse, often revealing only task accuracy and (indirectly) any latent attributes that impact their behavior. This places corresponding limits on the kinds of inferences that people can make about these agents.

Finally, recent work on cognitive models of *pedagogy* has also wrestled with the question of how we represent another agent's knowledge or capabilities (especially when they differ from our own). This work emphasizes the idea that when conveying knowledge, teachers reason about what would be most helpful for learners and learners maintain a similar model of the teacher that reflects their intention to provide helpful information (Aboody et al., 2023; Bass et al., 2022; Bass et al., 2019; Bonawitz et al., 2011; Gweon, 2021; Shafto et al., 2014). While such mental models of teacher and learner appear critical to accounting for human social learning even in simple tasks, it is less clear how evaluations of another agent's knowledge impact collaborative behavior, particularly over repeated interactions in which knowledge may change.

Taken together, recent work across psychology and artificial intelligence points to people's sophisticated ability to infer the competence of those around them and integrate these inferences into their decision-making in collaborative settings. However, these findings also suggest that much of the richness of our internal models of others remains unaccounted for in existing tasks requiring evaluations of other agents. First, in everyday settings, people's abilities are rarely static. In the monkey bars example at the outset, a parent's decision today will likely have to be re-evaluated in weeks or months. The notion of others as *learners* that is central to work in pedagogy should be incorporated into our representations of others' abilities in other collaborative settings—how fast collaborators learn and what it is they're learning may be just as important as their competence itself. Second, people's abilities in everyday settings are often rich and multi-layered. A person is not merely a "good" piano player or scientific writer; rather, our representation of their ability reflects a decomposition of the task into relevant skills and knowledge that people may express to varying degrees. Thus, our evaluations of what others are capable of in collaborative settings should reflect the richness of social inference more broadly (Gweon, 2021; Vélez & Gweon, 2021).

The current work aims to address both of these challenges. First, rather than collaborating with an agent whose ability is static, we explore people's behavior over many interactions with agents whose competence changes over the course of the task. Though far from reflecting the full richness of human learning, this work allows us to ask how much another agent's *improvement* plays a role in our representations of them and in subsequent collaborative decision-making. Second, in contrast to prior work which has explored collaborative behavior in fairly abstract domains such as games or stock investing (Aboody et al., 2023; Leong & Zaki, 2018; Vélez & Gweon, 2019), the current work examines behavior in a physics-based setting where people's decisions draw on a rich internal model of the task's underlying mechanics. As a result, collaborators' actions can reflect knowledge structures that differ from participants' own in systematic and subtle ways and, critically, participants' own mental model may allow them to form more nuanced ability judgments on the basis of their partner's errors. In this way, we aim to align the diverse research threads summarized above towards a more comprehensive account of how we represent others' competence in collaborative settings.

We investigate people's ability to collaborate with an artificial agent across many repeated interactions in a physics-based video game. Participants were tasked with catching a ball launched from different locations around a circle by placing a paddle where the ball would land. Each round, they were given a suggestion from their agent partner about where to place the paddle to catch the ball. Building on prior work, we use people's decisions about whether to accept or modify their partner's suggestions to probe underlying representations of the partner's competence (Chen et al., 2020; Xie et al., 2019). In experiment 1, people's agent partners differed in their true competence over time, parameterized here as the variance of their suggestions. We first ask how much people's behavior in the game draws on their own physical judgments versus the suggestions of their partner and how this varies based on their partner's competence. Next, we ask whether people's intervention decisions reflect a dynamic representation of their partner's ability above and beyond the trial-specific context. In experiment 2, we incorporate a richer notion of the agents' *competence* by modifying their underlying representation of the task environment. In two conditions where the agents' beliefs about the mass of the ball differ from the ground truth, their paddle suggestions exhibit systematic errors that are diagnostic of this inaccurate world model. We investigate participants' ability to correct for these errors and, critically, whether they can do so using only the structure of the errors themselves rather than their own mental model of the task.

Our results contain several key findings: First, rather than relying exclusively on their own physical judgments or the advice of their partner, people integrated both sources of information in their interventions. Moreover, the degree to which they incorporated their partner's input was predicted by how reliable the agent had been in the past, not just the quality of its current advice. In experiment 2, we find that people can accurately correct for their partner's biased suggestions and further, can integrate the underlying structure of their partner's error into intervention decisions in the absence of other cues. Taken together, our results suggest that across repeated interactions, people's collaborative behavior reflects dynamic inferences about others' abilities and even their internal models of the task itself.

3.2 Experiment 1

3.2.1 Participants

256 adults recruited from Prolific completed the task online. Data from 12 participants were excluded from subsequent analyses due to technical issues encountered during the experiment, resulting in 244 participants with complete data (average age: 33.8 years, SD = 11.3; 127 male, 103 female, 13 non-binary; educational background distributed across high school, 4-year college, and graduate degrees). The experiment was designed to last roughly 25 minutes (average completion time: 23.3 minutes, SD = 9.0) and participants were paid $14/hr based on this expected completion time. All participants provided informed consent in accordance with the UC San Diego IRB.[1]

3.2.2 Human-agent collaboration task

In the experiment, participants tried to catch a virtual ball launched from a point on a circle using a rectangular paddle positioned along the outside of the circle (see Figure 3.1). Participants worked together with an artificial agent "partner" who was

[1] All code used to run the experiment, as well as code used in analyses below, can be found at: https://github.com/erik-brockbank/physics_agent_manuscript.

trying to help them on the task. On each round, the partner suggested a paddle location based on the ball's launch position; participants could either accept this suggestion or adjust the paddle themselves before launching the ball.

Each trial began with participants' agent partner suggesting a paddle location that would catch the ball; the paddle was shown moving around the circle and a small animation on the right showed the agent "thinking." Once the agent had moved the paddle to its suggested location, participants were given the opportunity to either adjust the paddle with the arrow keys or keep their partner's suggestion. If participants adjusted the paddle, the agent's original recommendation remained visible and marked in gray. When participants settled on a paddle location, they launched the ball with the spacebar. The ball's path was animated and participants were shown a message indicating whether they had successfully caught it before proceeding to the next trial. Every session consisted of 96 trials divided into eight blocks of 12. This "block" structure was not visible to participants; in each block, the ball appeared at locations sampled in a random order from each of 12 bins of equal width along the circle's circumference. The 96 launch locations were determined before the experiment and were identical for all participants.

3.2.3 Manipulating agent ability

Participants were assigned to one of four conditions that manipulated the quality of their partner's suggested paddle locations: a *reliable* partner, an *unreliable* partner, an *improving* partner, and a *worsening* partner. The agent's suggested paddle location on each trial was an angle θ sampled from a *von Mises* distribution (a circular approximation to a normal distribution) with mean μ equal to the ball's final landing angle ρ, and variance σ^2 determined by the agent's competence level. The *reliable* agent had a low $\sigma^2 \approx 10$ degrees, ensuring that the sampled paddle suggestion was almost always close to the ball's true landing location. By contrast, the *unreliable* agent sampled its paddle locations from a high-variance distribution with $\sigma^2 \approx 48$ degrees. The high and low-competence σ^2 values were chosen to give the agents expected success rates of around 80% and 20%, respectively. Meanwhile, the *improving* agent began with a σ^2 value equal to the *unreliable* agent's but every 12 trials the variance decreased by a fixed amount so that during the final 12 trials, it had a σ^2 equal to the *reliable* agent's. The *worsening* agent showed symmetrical changes to its σ^2 but in the opposite direction, beginning like the *reliable* agent and performing like the *unreliable* agent in the final block.

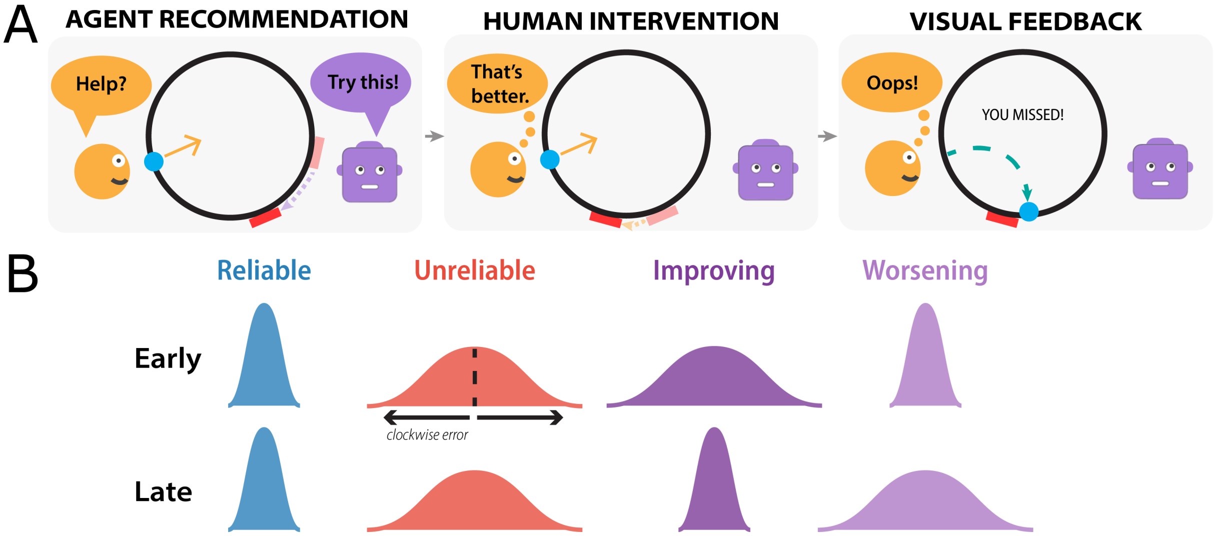


Figure 3.1. Task and Experiment 1 Design. (A) Participants worked with an artificial agent partner to catch a ball launched from the edge of a circle. Their partner began by suggesting a paddle location which participants could either accept or modify. (B) The agents chose suggested paddle locations from a distribution around the ball's true landing position. The variance of this distribution determined how reliable the agent's suggestions were. Participants were assigned to one of four conditions that varied the reliability of the agent's paddle suggestions over the course of the experiment.

3.2.4 Measuring human appraisals of agent ability

A core goal of our study was to investigate the impact of manipulating an agent's behavior on participants' impressions of its competence, thereby impacting how they approached collaborating with it. We measured participants' appraisals of their partner's task ability as the degree to which they intervened before committing to a final paddle location on each trial. Intuitively, participants who judged their partner to be more competent would be less likely to revise their partner's suggestion, or do so to a lesser extent. On each trial, we measured whether participants intervened to adjust the paddle's position away from their partner's initial suggestion and the magnitude of this intervention.

If participants were maintaining an ongoing estimate of their partner's task competence, their intervention behavior might be guided by this estimate above and beyond the trial-specific accuracy of their partner's suggestions. For example, participants might place more confidence in the suggestions of the *reliable* agent relative to the *unreliable* agent, even when the magnitude of the error in the agents' recommendations are equal. To isolate the impact of learned expectations about each agent's ability on participants' interventions, we included a *critical trial* in each 12-trial block (unbeknownst to participants): Rather than sampling locations as described above, the suggested paddle location on critical trials was set to a fixed distance from the ball's landing location (approximately 16 degrees) that was close to the true landing angle yet would result in missing the ball unless the participant intervened. Including these critical trials enabled direct comparisons between conditions while controlling for the magnitude of the error in the agent's suggestion.

3.2.5 Post-study questionnaire

After completing all 96 trials, participants were given a post-study questionnaire to collect basic demographic information described above as well as two additional demographic variables we did not analyze here: prior physics courses taken and prior experience with video games. Next, they were asked how often they thought they had intervened on the previous trials and how often they would expect to intervene if they were to play another 96 rounds with this same partner (both 1-100% scales). Finally, they were asked to indicate how much they trusted the agent to catch the ball (five-point rating scale) and to describe how they decided whether to intervene in the task. We do not analyze these data in the current results but the full set of responses are available along with the trial data and experiment code.

3.3 Results

We begin by examining the performance of human-agent teams on the task overall. They caught the ball on 73.8% (SD = 14.7%) of trials across all conditions, improving from 55.9% in the first trial block to 82.8% in the final block. The root mean squared error (RMSE) of the final paddle locations was 14.85 degrees (SD = 7.56 degrees). Together, these findings suggest that while the task was challenging at the outset, participants were nevertheless able to achieve reasonably high performance. However, our primary interest is in how their behavior varied across conditions as a result of differences in their *bot partner's* ability. In the analyses below, we first ask what role the bot partner's suggestions played (if any) in participants' final paddle placements: How much did participants incorporate their partner's suggestions into their decisions about where to position the paddle? Next, we explore whether participants' collaborative behavior also reflected broader inferences about their partner's ability; in other words, did their paddle placements rely not only on the bot's current suggestion, but on the accuracy of its prior suggestions?

3.3.1 People combined information sources to make intervention decisions

To understand how participants coordinated with their partner, we compare three possible accounts: First, it may be that people trusted their bot partner *completely*, regardless of its competence. On this view, participants' own physical intuitions would have played no role in their decisions. A second account takes the opposite perspective; people may have *ignored* their partner's suggestions entirely, simply choosing the best paddle position each round without regard for the bot's initial proposal. Finally, a third possibility is that people's behavior was somewhere in the middle. Rather than consistently following their partner's suggestion or unilaterally seeking the optimal paddle position each round, people may have relied on a *combination* of their own physical intuitions and their partner's recommendation to decide where to place the paddle. We consider each of these options below; our results suggest that participants integrated intuitive physical judgments with their partner's guidance and that *how much* they incorporated their partner's suggestions was calibrated to their partner's task performance.

Participants intervened to improve accuracy

We start by considering the first hypothesis above, that people merely acted in accordance with their partner's suggestions. If this were true, we would expect intervention rates to be low and performance in each condition to closely match the ability of the agents in that condition. Figure 3.2 (Top) shows average intervention rates (the percent of trials in which each subject modified the agent's original suggestion) in each trial block. Notably, intervention rates were high in all conditions, even with the *reliable* agent, whose suggestions would catch the ball on approximately 80% of trials. Figure 3.2 shows an overall increase in intervention rates even in the *reliable* and *unreliable* conditions where agent performance did not change. This seems most likely to be a result of participants' overall task improvement noted above. A generalized linear mixed effects model fit to participants' intervention decisions (binary) with a random intercept and slope for each participant showed that intervention rates differed significantly across trial blocks ($\chi^2(1) = 40.76$, $p < .001$); further, there was a significant interaction between trial block and condition ($\chi^2(3) = 97.96$, $p < .001$), indicating that changes in intervention rate over time differed significantly between conditions. Thus, far from merely trusting their partner's suggestions, participants took an active role in intervening and calibrated their interventions to their partner's underlying ability.

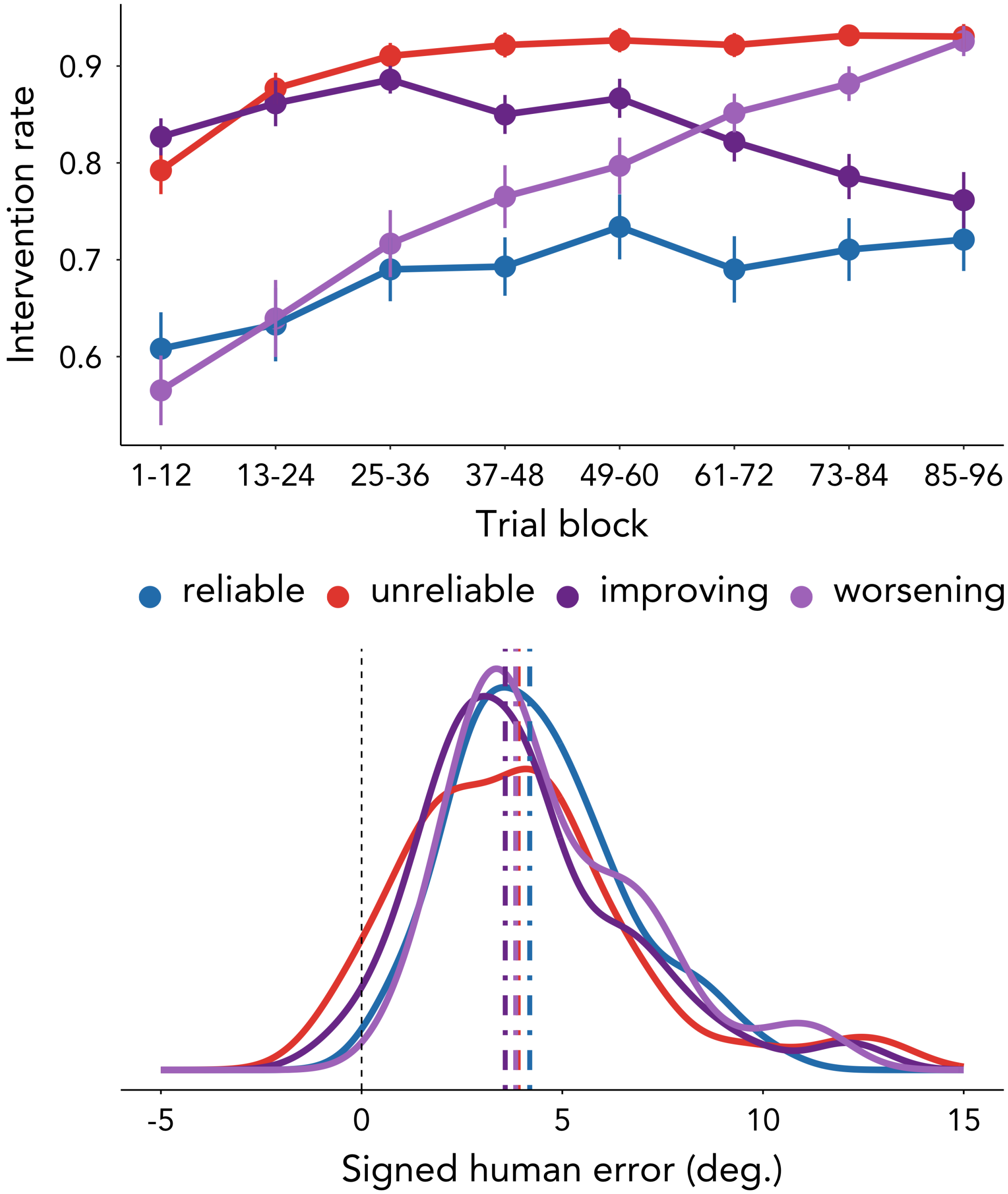


Figure 3.2. Performance with each agent partner. (Top) Paddle intervention rates by trial block. Error bars show standard error of participant means. (Bottom) Distribution of average participant error by condition. Positive values indicate responses whose offset from the correct paddle location were in the same direction as the agent's suggestion and negative values indicate the opposite. Curves are based on kernel density estimation. Colored vertical lines indicate medians in each condition.

Intervention decisions incorporated agent suggestions

In light of the high intervention rates across conditions, one account of people's behavior is that they simply relied on their own intuitive physics to respond, moving the paddle to their best guess about the ball's landing place regardless of where the agent first positioned it. On this view, the quality of their partner's recommendations would have been irrelevant. To test this possibility, we examine participants' *errors* on each trial—the distance (in degrees) of the paddle from the ball's final landing location. Intuitively, if people completely disregarded their partner's suggestions, the distribution of their errors would be centered on the ball's true landing location. Alternatively, if people took their partner's suggestions into account, we might expect final paddle placements to be systematically biased towards or away from the partner's initial suggestion. Figure 3.2 (Bottom) shows the distributions of participants' average error in each condition. Critically, these distributions are signed relative to the agent's paddle suggestion; error greater than 0 represents participants placing the paddle away from the ideal catching location *in the direction of the agent's suggestion.* Meanwhile, error less than 0 represents participants placing the paddle away from the ideal location *in the opposite direction of the agent's suggestion.* Participants' average signed error was significantly greater than 0 in all four conditions, reflecting a stable bias toward their partner's recommended paddle locations (*reliable*: $t(56) = 15.70$; *unreliable*: $t(66) = 8.13$; *improving*: $t(64) = 12.34$; *worsening*: $t(54) = 14.20$; all ps $< .001$). A linear mixed effects model fit to participants' signed error on *individual* trials with random intercepts and slopes for each participant showed similar results: Signed error distributions were not significantly different between conditions ($\chi^2(3) = 6.06$, $p = .11$) but 95% confidence intervals on the estimated marginal means were greater than 0 in all conditions (*reliable*: mean = 4.04 degrees, 95% CI = [3.29, 4.80]; *unreliable*: mean = 3.28, 95% CI = [2.59, 3.98]; *improving*: mean = 3.09, 95% CI = [2.39, 3.79]; *worsening*: mean = 4.20, 95% CI = [3.44, 4.97]). This suggests that people's decisions about where to place the paddle were not merely an effort to find the best location independent of the bot's suggestion; rather, they showed a systematic anchoring toward their partner's recommendation.

Taken together, the results in Figure 3.2 suggest that people's decisions about where to place the paddle integrated multiple sources of information. They did not naïvely trust their agent partner regardless of its competence, nor did they simply choose the best move each round without consideration for their partner's recommendation. However, the agent's suggestion on a given trial is not the only source of information that might help participants decide where to ultimately place the paddle. Across repeated interactions, agents in each condition offer ongoing evidence of their underlying *competence* through the accuracy of their paddle suggestions. Participants can use this information to calibrate how much their final paddle locations should be influenced by their partner.

3.3.2 People relied on past performance to guide interventions

Since agent partners varied across conditions in how helpful their paddle suggestions were, we hypothesize that participants incorporated this information into their decisions about how closely to follow their partner's suggestions. To test this, we begin by looking at the relationship between the agent's paddle suggestion error and participants' paddle intervention magnitude across conditions. If participants were correcting for the agent's errors in a way that did not integrate the agent's underlying ability, this relationship should be similar across conditions (i.e., they should adjust for small errors less and larger errors more in a similar fashion). We fit a linear mixed effects model of participant intervention magnitude (on trials in which they intervened) as a function of agent recommendation error and condition with a random intercept and slope for each participant. We find a significant interaction between condition and agent error ($\chi^2(3) = 351.9$, $p < 0.001$), suggesting that the agent's competence played a critical role in the way people's *intervention* magnitudes varied with the bot's initial *suggestion error*. However, this result could be driven in part by the fact that the underlying distribution of agent errors differed substantially across conditions (by design). Thus, a more apples-to-apples comparison should examine people's intervention behavior for similar levels of agent error across conditions. For this, we turn to the eight *critical trials* that each participant completed.

Critical trial interventions reflected differences in agent ability

If people's responses combined their own estimate of the ball's final location and their partner's suggestion—without considering their partner's overall reliability—we should not see any difference in intervention behavior on the critical trials, since the agent's paddle suggestion error on critical trials was the same across conditions. Figure 3.3 (Top) shows average intervention rates on critical trials. Though participants intervened less than they should have across the board, there is a clear difference between conditions that aligns with differences in the bot partner's competence: Participants intervened more often with an *unreliable* or *improving* partner than they did with a *reliable* or *worsening* partner. To quantify these differences, we fit a generalized linear mixed effects model to participants' interventions (binary) on critical trials with a random intercept for each participant; intervention rates differed significantly across conditions ($\chi^2(3) = 17.84$, $p < 0.001$). Estimated marginal mean intervention rates were similar to the averages shown in Figure 3.3 (*reliable*: mean = 79.41%, 95% CI = [73.01%, 84.55%]; *unreliable*: mean = 89.28%, 95% CI = [85.16%, 92.27%]; *improving*: mean = 90.38%, 95% CI = [86.50%, 93.28%]; *worsening*: mean = 81.61%, 95% CI = [75.55%, 86.41%]) and differed significantly between *unreliable* and *reliable* conditions ($p = .01$), as well as *improving* and *reliable* ($p = .003$) and *improving* and *worsening* ($p = .021$). Thus, decisions about *when* to intervene on critical trials were sensitive to differences in the agents' abilities.

Participants' decisions about *how much* to intervene on critical trials (Figure 3.3, Bottom) shows a similar pattern; those paired with an *unreliable* or *improving* partner made larger adjustments than participants whose partner was highly accurate (*reliable*)

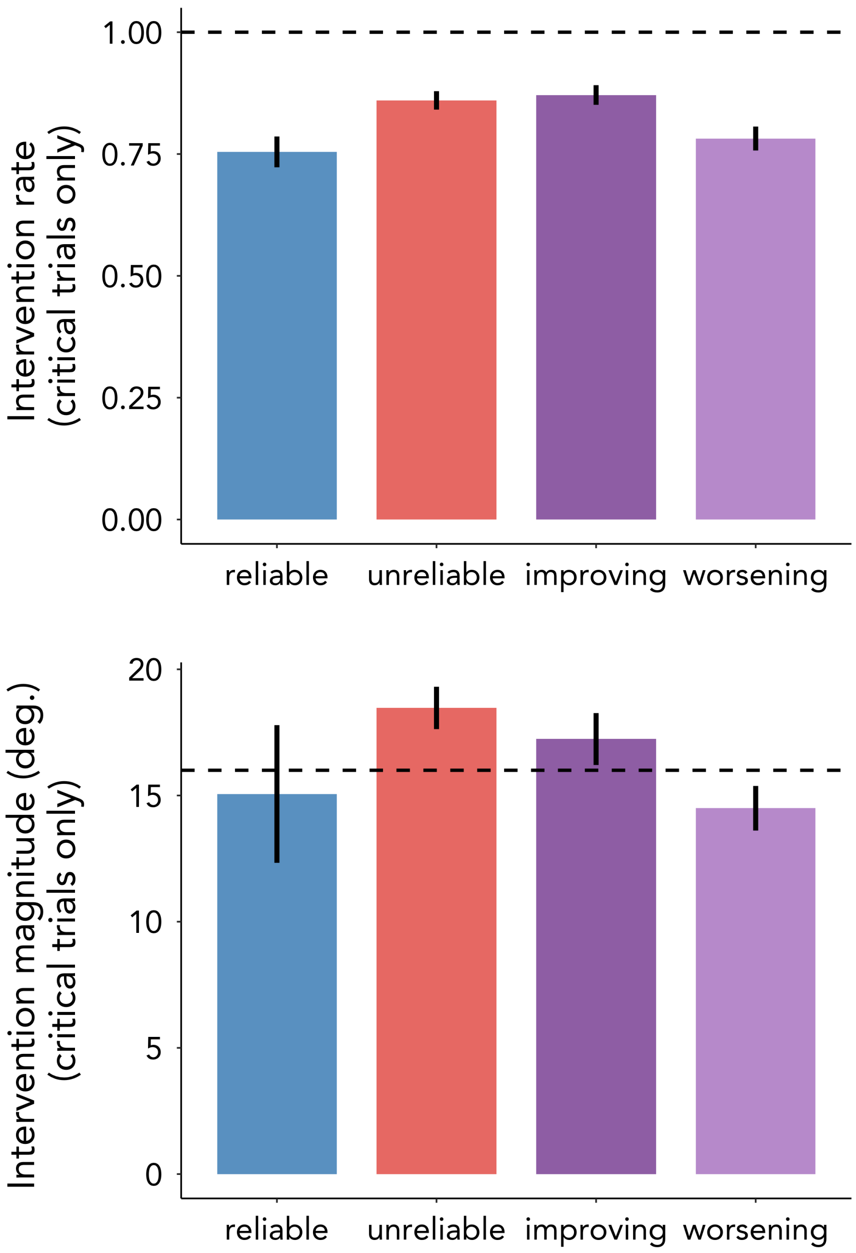


Figure 3.3. Intervention behavior on critical trials. (Top) Proportion of critical trials on which participants chose to intervene. The dashed line indicates optimal behavior (critical trials always required intervention to catch the ball). (Bottom) Distance participants intervened on critical trials. The dashed line indicates the optimal intervention distance on these trials. Error bars in both plots represent standard error of participant means.

or started out accurate (*worsening*). We fit a linear mixed effects model to participants' intervention magnitude with a random intercept for each participant, now looking only at critical trials in which participants intervened. We found significant differences between conditions in intervention distance on these trials ($\chi^2(3) = 28.33$, $p < .001$). As with intervention rates, estimated marginal mean intervention magnitudes were similar to the participant averages shown in Figure 3.3 (*reliable*: mean = 12.1 degrees, 95% CI = [10.4, 13.9]; *unreliable*: mean = 18.3, 95% CI = [16.7, 19.8]; *improving*: mean = 17.0, 95% CI = [15.4, 18.6]; *worsening*: mean = 14.2, 95% CI = [12.5, 16.0]). These showed significant differences between *reliable* and *unreliable* agents ($p < .001$), *reliable* and *improving* agents ($p < .001$), and *unreliable* and *worsening* agents ($p = .005$). Thus, a complete account of reasoning on this task suggests that people maintain an underlying assessment of their partner's competence over time and calibrate their decisions about whether to intervene, and how much, based on this assessment.

3.4 Discussion

In this study, we address the question of how people evaluate an artificial agent's competence in a collaborative physical prediction task. Specifically, we investigated how differences in an agent's ability impacted people's decisions to either trust their partner's recommendation or intervene to modify it. Our results contain two key findings. First, we show that participants integrate their own physical judgments and the recommendations of their partner, intervening frequently in a way that was often "anchored" toward their partner's suggestion. Second, we find that on trials in which the agents' suggestions were identical across conditions, participants differed significantly in the frequency and magnitude of their interventions in a way that reflected their partner's prior success. This suggests that the process by which people integrate their partner's recommendation involves a dynamic inference about their partner's *ability*; people calibrate how much to defer to their partner based on the prior reliability of their partner's suggestions and are sensitive to changes in their partner's performance over time.

What kind of representation of their partner's competence did participants have? In the current experiment, each bot's ability could be distilled to a single parameter which either varied over time or was constant. As a result, participants did not need to form a sophisticated representation of their partner's competence; it would have been sufficient to maintain a simple dynamic estimate of the bot's accuracy or the variance of their suggestions. Yet in the real world, estimates of another's abilities are far more nuanced. Consider the reasoning a teacher must do about their students' knowledge to understand why they made certain errors on their homework, or the way a soccer coach might evaluate a player's behavior in games to identify areas of improvement. In these settings, evaluations of competence extend far beyond a single parameter related to task accuracy. Instead, they reflect rich inferences about another person's *internal model* of the task—what they know or are capable of and how they make decisions. How do people develop these complex representations of others' competence over repeated interactions? To better understand this, we extend our first experiment to include agents whose fundamental knowledge about the physical dynamics of the task differs from the ground truth. As a result, these agents produce systematic patterns in their errors which reflect their misaligned internal models. We once again measure participants' intervention behavior, now with the aim of better understanding how people represent another agent's model of the world and use this representation to support collaboration.

3.5 Experiment 2

In this experiment, we investigate people's ability to collaborate with an agent whose understanding of the physical dynamics of the environment differs systematically from the ground truth. This represents a broader and more naturalistic inquiry into how people evaluate the competence of others. In experiment 1, we find evidence that participants' collaborative behavior—their decisions about when to intervene on their partner's suggestion, and how much—reflects an ongoing evaluation of their partner's *underlying ability* beyond mere trial-by-trial decision making. However, these results license limited inferences about how richly participants represent their bot partner's competence. This is because the bot agents in the previous experiment differ only in the variance of their suggestions around the ball's true landing location. Therefore, on any given trial, the bot agent's competence can be captured by this single latent parameter. In this way, the challenge for participants of inferring their partner's ability in the previous experiment could be as simple as updating this parameter estimate after each trial. This task of estimating a latent parameter in the environment is interesting in its own right, and is likely something we do often (consider for example grabbing a jacket on the way out of the house based on the inference that it often rains this time of year). However, when evaluating the competence of others, people are also capable of far richer representations of another agent's ability. Such representations serve a useful purpose: they allow for better prediction and may generalize to other similar tasks. How people make such complex inferences about others' competence in service of collaborative goals remains an open question. The current experiment approaches this broader question by requiring participants to make more sophisticated inferences about their bot partner's ability. We retained much of the structure of the previous experiment, while making two key modifications.

First, the competence of the bot agents was manipulated through their knowledge about the physical dynamics of the task itself, which led to systematic errors (in contrast, the agents in experiment 1 had normally distributed errors which didn't result from differing knowledge about the task). Concretely, bot partners in the current experiment had distinct beliefs about the *mass* of the ball being launched. In two conditions, these mass estimates were higher or lower than the true mass of the ball; as a result, these bots consistently over- or under-estimated the path of the ball, and the *amount* by which they did so varied depending on its launch location. Thus, while participants in the first experiment could not extract anything from their bot partner's errors other than a refined estimate of the bot's accuracy, participants in this experiment faced a more nuanced inference problem. The bot's suggestions revealed information about the bot's *internal model* of the task environment—namely its belief about the weight of the ball—which could explain the errors it made. Participants who inferred this about their partner could use such a representation to systematically correct their partner's suggestions.

We are interested in whether participants' intervention behavior reflected an understanding of this predictable structure in their partner's errors. In the previous experiment, we probed participants' social inferences about their partner's underlying ability, and the role that these inferences played in their intervention behavior, using *critical trials* in which bot error was equated across conditions. The second major change in the current experiment is to extend those results by exploring whether participants inferred an accurate *predictive model* of their partner's paddle suggestions. In eight *mystery round* trials throughout the experiment, the ball's launch location was hidden from participants using an occluder. Participants were told that their bot partner could see the ball's true launch location, but participants saw only the bot's paddle suggestion. In the absence of information about where the ball was being launched from, participants could not adjust the paddle using their own inferences about the ball's path. Instead, they needed to decide whether to adjust their partner's paddle placement, and how much, based only on the bot's own suggestion. Here, an understanding of *how the bot generates its suggestions* could allow participants to intervene in the right direction, and roughly the right magnitude, just based on the suggested paddle location. We assessed participants' intervention behavior on both normal trials and mystery round trials to understand how participants collaborated with a partner whose knowledge differed from their own, and how well they inferred this difference in knowledge to further support collaboration.

3.5.1 Participants

180 adults were recruited from Prolific; two were excluded due to technical issues, leaving 178 participants with complete data. Participants were native English speakers from the United States and United Kingdom who had not participated in the previous experiment. Average age was 40.7 years ($SD = 14.08$); 78 participants were female, 95 were male, and 2 were non-binary. Education spanned high school (49), two- and four-year college (29 and 36), graduate school (36) and post-graduate (25). We anticipated the experiment would last roughly 30 minutes and participants were paid $15.50/hr based on this expected completion time (average completion time: 22.3 minutes, $SD = 6.48$). All participants provided informed consent in accordance with the UC San Diego IRB.[2]

To determine our sample size, we ran a power analysis based on experiment simulations that surfaced estimates for a key measure in our analyses. We simulated N participants each completing eight mystery round trials with a true probability P of moving the paddle in a direction that corrects for the bot's underlying bias. We then ran a one-sided t-test of the hypothesis that the true probability of correcting the bot's bias is greater than chance (50%) using the proportion of trials with a response in the right direction for each of our N simulated participants. Based on 10,000 simulated experiments repeating this process, we estimated that with 50 people in a condition, we would have greater than 99% power to detect a true probability of correcting the bot's bias of $P = 0.6$. A probability of 0.6 represents a "small" effect size (Cohen's $h = 0.2$) relative to a chance probability of 0.5 (Cohen, 1977). Therefore, we aimed for 50 people in each condition for a total of at least $N = 150$ participants across our three conditions. We recruited 180 in total to allow for the possibility of technical error and imbalances during condition assignment.

[2] All code used to run the experiment, as well as code used in analyses below, can be found at: https://github.com/erik-brockbank/physics_agent_manuscript.

3.5.2 Task overview

The current experiment was structured nearly identically to experiment 1; participants completed 96 trials divided into eight blocks of 12 (participants were not informed of this structure, only the total number of trials). In each trial, participants were shown the ball's launch location and angle, along with an animation of the bot's paddle suggestion. They were then given a chance to modify the bot's suggestion before launching the ball. After launching the ball, its path was animated and participants saw whether their paddle placement was successful (see Figure 3.4A).

In the current experiment, participants also completed eight *mystery round* trials (Figure 3.4B), one in the second half of each trial block. This was done to ensure that the first mystery round occurred after participants had received some exposure to their bot partner's suggestions and that all subsequent mystery rounds were roughly equally spaced throughout the experiment. The mystery round launch angles were selected from eight of the 12 launch angle "bins" used in each trial block. The eight launch locations were identical for every participant. The order of the eight mystery rounds was shuffled for each participant and assigned to a random trial in the second half of each trial block. Mystery round trials proceeded nearly identically to normal trials. At the beginning of the trial, participants were shown a large quarter-circle occluder hiding the quadrant of the circle from which the ball was being launched. They next saw their bot partner suggest a paddle placement as in other trials; after deciding whether to revise the bot's suggestion and launching the ball, the occluder was removed to show the ball's launch location for a brief interval before animating the ball's path and whether participants successfully caught it.

3.5.3 Manipulating agent ability

In the previous experiment, participants were paired with one of four bot partners whose competence was reflected in the *variance* of their paddle suggestions around the

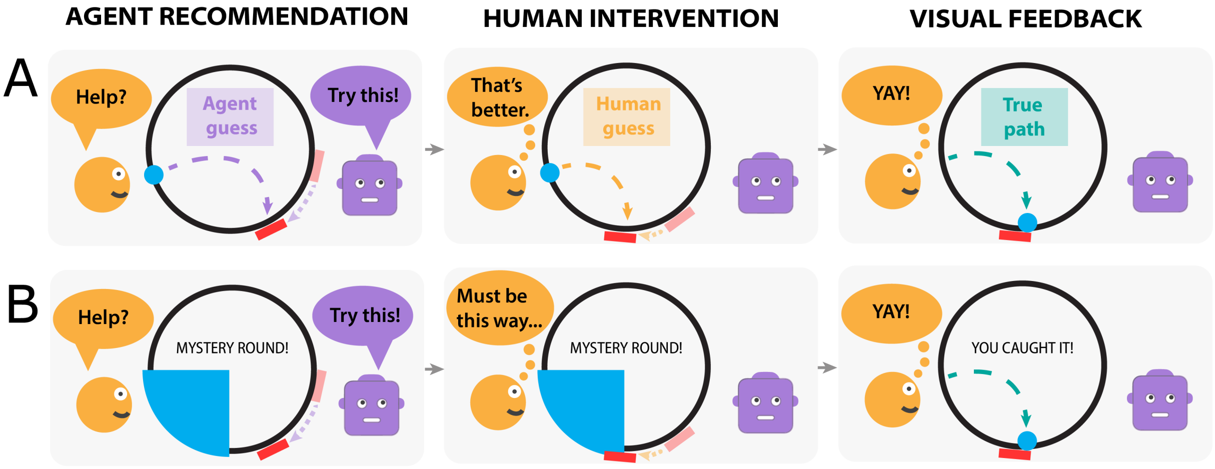


Figure 3.4. Experiment 2 Overview. (A) Participants' agent partners had a biased estimate of the ball's path which systematically influenced their paddle suggestions. (B) On "mystery round" trials, the ball's launch point was hidden from participants. They were told that their partner could still see the launch location; their job was to adjust the paddle using only their partner's suggestion as a cue.

ball's true landing location (Figure 3.1). In this experiment, we instead manipulated bot ability through the *bias* in their suggestions. Participants were assigned to one of three conditions. In all three conditions, the bot sampled its paddle suggestion each round from a *von Mises* distribution centered at the angle $\hat{\rho}$ where the bot *estimated* the ball would land on that trial. The variance σ^2 of this distribution was identical in all conditions and was chosen so that roughly 90% of paddle suggestion angles sampled from this distribution would in fact catch a ball landing at its center. Thus, all three bots provided reliable paddle suggestions based on their estimate of the ball's landing location. However, they varied in the way they estimated where the ball would land.

In the *heavy bias* condition, the bot's estimate of the ball's mass was 50% larger than its true mass. As a result, the bot consistently *underestimated* the path of the ball and its corresponding landing location. Critically, the magnitude of these underestimations varied depending on the ball's launch location around the circle. For launch angles close to the left and right center points (3 and 9 on a clock face), where the ball had the largest horizontal distance to travel, the bias was most evident. Meanwhile, for launch angles close to the top and bottom center points (6 and 12 on a clock face), the bias was almost non-existent, since an incorrect estimate of the ball's mass does not impact where it lands when launched straight up or down. This means that for participants in this condition, correcting the bot's suggestions was not a matter of learning a fixed adjustment magnitude that would correct their partner's bias in all cases, but rather a continuous mapping between the launch angle and the bot's corresponding underestimation.

The *light bias* condition was identical to the *heavy bias* condition, except that the bot's belief about the ball's mass led it to consistently *overestimate* the path of the ball through the circle. Concretely, this bot's estimate of the ball's mass was 50% smaller than the ball's true mass; in this condition, we also modified the bot's estimate of the ball's *launch velocity* so that the magnitude of the bot's overestimates matched the underestimates in the *heavy bias* condition. In short, the *light bias* bot made symmetrical errors to the *heavy bias* bot, but in the opposite direction.

Finally, in the *no bias* condition, the bot's estimate of the mass of the ball reflected the ground truth. Its paddle suggestions on each round were sampled from a von Mises distribution centered at the ball's *true landing location* rather than a biased estimate of the ball's landing location. This condition provides a control where the agent's suggestions lacked a systematic bias; instead, its error pattern can be thought of as resulting from noise, similar to the agents in the previous experiment. We explore the way people infer the competence of an agent who has a fundamentally different internal model of the world in the two bias conditions, versus one who understands the environment similarly but may sometimes act differently due to random error.

3.5.4 Measuring representations of agent ability

As in the previous experiment, our analyses focus on how participants intervened on their bot partner's paddle suggestions. On each trial, we recorded whether participants intervened, the magnitude of their interventions, and the corresponding error of the bot's original suggestion as well as participants' final paddle placements. We hypothesized that the frequency and magnitude of participants' interventions reflected not merely their ability to collaborate with the bot on the task, but their underlying inferences about their partner's *competence* over the course of the experiment. Comparing these values across conditions allows us to assess how participants' appraisals of their bot partner's abilities played out in their collaborative behavior.

However, an important question in our results is how much participants' behavior truly reflected an underlying inference about their bot partner. Our results seek to disentangle participants' judgments about their partner from their own intuitions about the task, which may develop independently of any reasoning about their partner's abilities. In the previous experiment, we compared responses on critical trials where the bots' suggestions were equated across conditions so that differences in behavior could be attributed to different judgments about the bot partners across conditions. *In the current experiment, we are interested in whether participants developed representations of their partner's ability which reflected their partner's internal model of the world*; this internal model differed from participants' own in the *heavy bias* and *light bias* conditions. To address this, we evaluate participants' behavior on the eight *mystery round* trials. As in the normal trials, we recorded whether participants intervened, how large their intervention was, and the magnitude of the resulting error along with the error of the bot's original suggestion. We hypothesized that participants' decision to intervene on these trials, and the magnitude and direction of these interventions, would reflect an understanding of the structured error patterns (or lack thereof) in their partner's suggestions.

3.6 Results

In our analyses, we explore two central hypotheses about people's collaboration with their bot partner. First, we predicted that participants would systematically correct for the bot's misaligned world model in the *heavy bias* and *light bias* conditions. We test this hypothesis in two ways. We begin by exploring participants' rate of intervention and overall accuracy between conditions to determine whether participants' intervention behavior reflected the differential abilities of their partner, and whether these interventions resulted in lower error. In addition, we explore the relationship between bot error and participant intervention magnitude *within* conditions to see whether participants were able to revise their bot partner's suggestions at a fine-grained level that was responsive to the variance in their partner's error.

The hypothesis above concerns participants' adaptive behavior, but does not address whether participants' interventions relied on a *representation* of their bot partner's world knowledge or the bias that this inaccurate knowledge produces. However, as noted previously, people's everyday inferences about the competence of others often involve reasoning about their internal model of a given task. We hypothesized that participants in the current experiment would show evidence of such inferences by being able to correctly adjust their partner's suggestions even when using *only the suggestions themselves* as cues. To test this, we examine intervention rate and error correction in the mystery round trials, where participants had visibility into the bot's suggestion but not the ball's launch location. Insofar as participants were able to systematically correct for their partner's bias without using the ball's launch angle to estimate its endpoint in the *light bias* and *heavy bias* conditions, this suggests participants had a richer representation of the *generative process* underlying their partner's suggestions.

3.6.1 People systematically corrected for their partner's incorrect world model

We begin by asking how well participants were able to correct their partner's errors, particularly in the *heavy bias* and *light bias* conditions. This question has two components: First, did differences in intervention behavior *between* conditions reflect the different demands placed on participants by the three bot partners? And second, did variation in intervention magnitudes *within* each condition align with variation in the bots' suggestion error? We find evidence in support of both questions, suggesting that broadly, participants were able to reliably and systematically correct for their partner's incorrect world model in the *heavy bias* and *light bias* conditions.

Participants intervened to improve accuracy

The bot partner in the *no bias* condition suggested paddle placements that would catch the ball without any changes in approximately 90% of trials. Participants therefore faced little pressure to modify the bot suggestions in this condition. In contrast, bot suggestions in the *heavy bias* and *light bias* conditions were frequently distant from the ball's true landing location—to succeed in these conditions, participants needed to intervene far more often and by larger amounts than in the *no bias* condition. Is this what they did? Figure 3.5 (Top) shows average *intervention rate*—the proportion of trials in which participants modified their partner's suggestion—in each condition over the course of the experiment. Intervention rates were relatively high in all three conditions and fairly consistent from beginning to end; however, there are also clear differences between the conditions. We fit a generalized linear mixed effects model to participant interventions (binary) on normal trials with random slopes and intercepts estimated for each participant. Consistent with the participant averages shown in Figure 3.5, estimates of marginal mean intervention rates were 92.27% in the *heavy bias* condition (95% CI = [88.69%, 94.78%]), 89.96% in the *light bias* condition (95% CI = [86.00%, 92.88%]), and only 48.10% in the *no bias* condition (95% CI = [39.12%, 57.13%]). Not surprisingly, intervention probability differed significantly across conditions ($\chi^2(2) = 72.43$, $p < .001$); pairwise comparisons of the estimated marginal means showed that intervention rates were significantly higher in the *heavy bias* and *light bias* conditions than in the *no bias* condition ($p < .001$ in both cases), but not different between *heavy bias* and *light bias* conditions ($p = .52$).

But how successful were these interventions? Figure 3.5 (Bottom) shows the average *error* of participants' paddle placement (i.e., the distance in degrees between the ball's true landing angle and the paddle's final angle) in each condition over the course of the eight trial blocks. In all three conditions, participants struggled more in the first block of trials but reached a fairly stable error level by the second block. Though a linear mixed effects model fit to participant error with a random intercept and slope estimate for each participant found significant pairwise differences between all three conditions (all ps <.001), here we are most concerned with whether participants achieved reasonable accuracy levels in each of the bot partner conditions. Estimated marginal mean error values from the mixed effects model were low across all three conditions, in line with the trend in Figure 3.5 (*heavy bias*: mean = 8.72 degrees, 95% CI = [7.97, 9.47]; *light bias*: mean = 6.80, 95% CI = [6.10, 7.50]; *no bias*: mean = 5.20, 95% CI = [4.47, 5.94]).

Graded intervention behavior was sensitive to partner bias

The results in Figure 3.5 suggest that between the three experimental conditions, participants were sensitive to differences in the overall magnitude of errors produced by their bot partner. However, beyond these differences between conditions, successful collaboration with each bot partner required correcting for *variation* in their suggestion error. Here, we ask whether participants were *calibrated* to their bot partner's errors by exploring the relationship between bot error magnitude and intervention magnitude within each condition. First, we examine whether participants consistently *improved* their bot partner's initial suggestion. How much participants improved their partner's

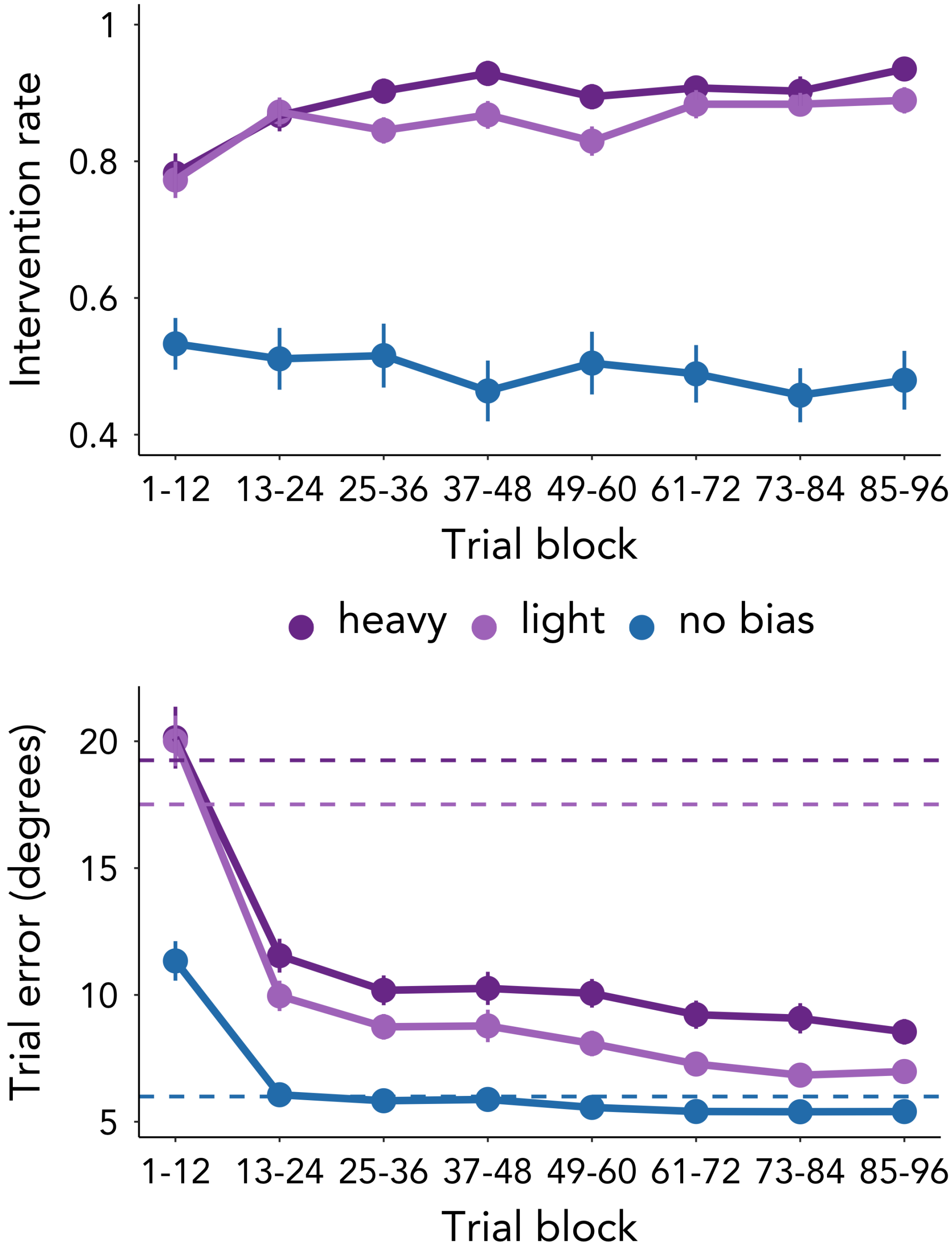


Figure 3.5. Behavior on normal trials by trial block. (Top) Paddle intervention rates. (Bottom) Error of participant paddle placements on normal trials. Dashed lines represent the average error of the bot's suggestions in each condition. Error bars in both plots represent standard error of participant means.

suggestion on a given trial is reflected in how close they came to perfectly correcting their partner's error. We quantify this as the difference between the error of the bot's original suggestion (its distance from the ball's true landing location) and the error of the participant's final paddle placement; we then scale this reduction in error by the error of the bot's original suggestion, resulting in a *proportion of bot error corrected* on that trial. This proportion has several intuitive properties. Values less than 0 result from participants making the bot's suggestion *worse*, either by moving the paddle in the wrong direction or by over-correcting more than the original error. Values of 0 signal that participants made no changes to the bot's suggestion or over-corrected by the exact same amount as the original suggestion. Finally, values between 0 and 1 arise from participants improving the bot's suggestion by some fraction of the original error, with better corrections closer to 1 (the proportion has an upper bound of 1 since the reduction in error achieved with the final paddle placement cannot be more than 100% of the original error magnitude).

Figure 3.6 (Top) shows the distribution of participants' average bot error correction proportions in each condition. Because the proportion of error corrected is vulnerable to large negative values in cases where the bot's error is small to begin with and participants intervene anyway, we remove trials in which the bot's error was in the bottom 20% of values before calculating these averages.[3] Participants in the *no bias* condition primarily made no change to their bot partner's error, with interventions relatively balanced in whether they improved or worsened the bot's suggestion. In contrast, the distribution of responses in the *heavy bias* and *light bias* conditions is shifted towards 1, reflecting consistent overall improvement to the bot's suggestions. To estimate the size of these differences, we fit a linear mixed effects model to individual participant bot error correction proportions with a random intercept and slope for each participant. When fitting the model, we once again excluded trials in which the bot's error was in the bottom 20% of values. We found a significant difference between conditions in the proportion of bot error corrected ($\chi^2(2) = 177.57$, $p < .001$); follow-up pairwise comparisons of estimated

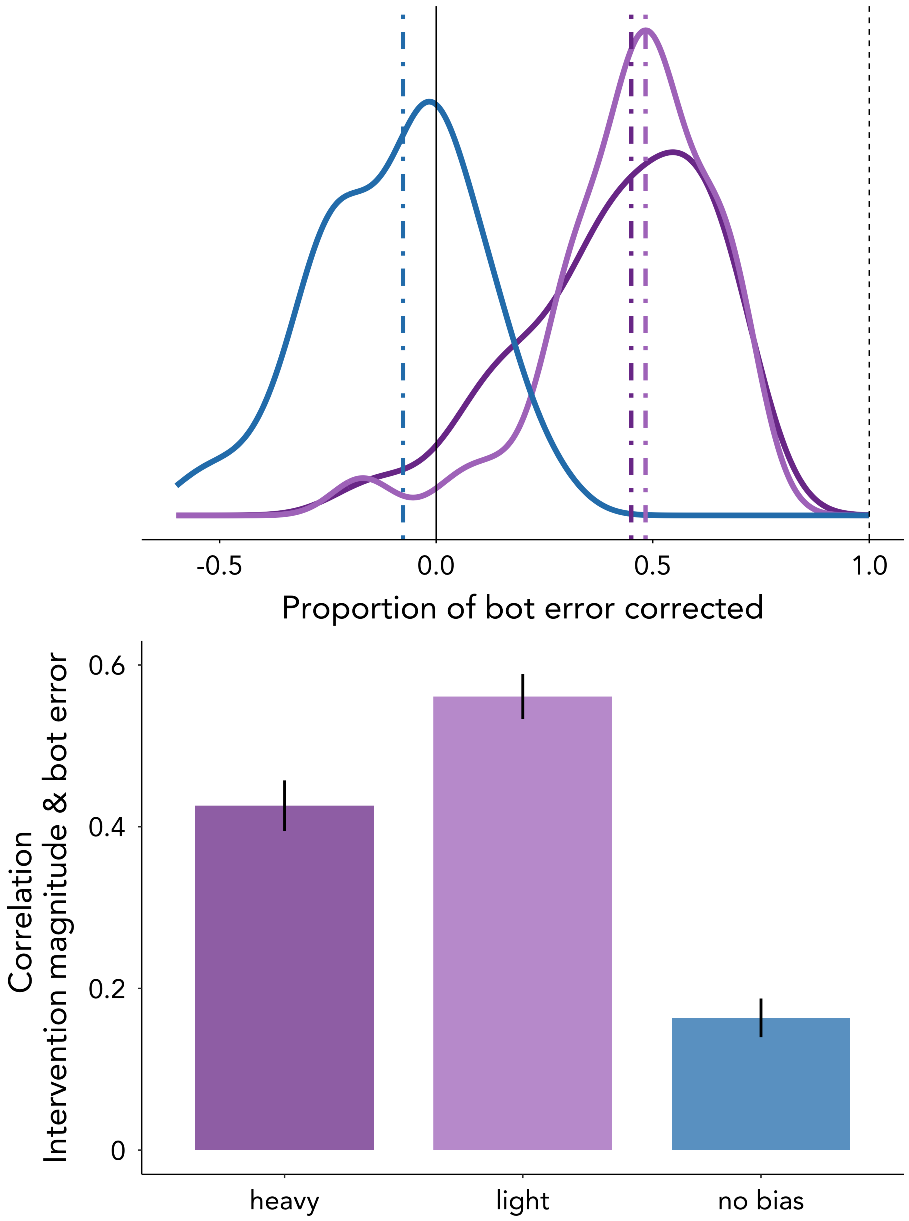


Figure 3.6. Participants' sensitivity to their partner's continuous error magnitudes. (Top) Distribution of average bot error correction proportions for normal trials in each condition after removing the bottom 20% of bot error trials. Values less than 0 result from modifying the bot's suggestion in the wrong direction; values of 0 represent no change in bot suggestion error, while values of 1 (theoretical maximum) represent complete correction of bot suggestion error. Vertical lines represent medians. (Bottom) Correlation between bot suggestion error (degrees) and magnitude of participant interventions for normal trials. Error bars represent standard error of participant correlations.

marginal means indicate that this is a result of significant differences between the *no bias* condition and the *heavy bias* and *light bias* conditions (both ps $< .001$) but not between the *heavy bias* and *light bias* conditions ($p = .67$). In line with the overall pattern in Figure 3.6, the estimated marginal mean error correction proportions are close to 0 in the *no bias* condition (mean = 2.37%, 95% CI = [-2.39%, 7.13%]) and are both similarly shifted towards 1 in the *heavy bias* (mean = 52.84%, 95% CI = [47.98%, 57.71%]) and *light bias* (mean = 55.76%, 95% CI = [51.22%, 60.31%]) conditions.

The bot error correction proportions in Figure 3.6 suggest that participants in the *heavy bias* and *light bias* conditions frequently corrected their partner's suggestions in the right *direction* and with varying magnitudes as a proportion of the bot's error. To better understand how closely participant intervention magnitudes scaled with the accuracy of their bot partner's suggestions, we compute the pairwise correlation for each participant between their partner's suggestion error and their intervention magnitude on all non-mystery round trials. Figure 3.6 (Bottom) shows the average of these individual correlations for participants in each condition. In the *heavy bias* condition, these correlations were significant for 75% participants; in the *light bias* condition, 89.1% were significant, and in the *no bias* condition, only 39.7% were significant. In line with this variation, a one-way ANOVA found that participant correlations differed significantly between the three conditions ($F(2, 241) = 52.47$, $p < .001$); follow-up t-tests confirmed that this was a result of significant pairwise differences between all three conditions ($p < .001$ for *heavy bias* and *light bias* versus *no bias*; $p = .002$ for *heavy bias* versus *light bias*).

To validate these results, we also fit a linear mixed effects model to participants' intervention magnitudes on each trial as a function of bot error magnitude and condition, with random intercepts and slopes for each participant (the linear relationship between participant intervention magnitude and bot error magnitude estimated by this model at

[3] Findings reported here are not highly sensitive to this choice of cutoff; smaller cutoffs primarily "penalize" error correction estimates in the *no bias* condition. Similar results are obtained when we instead set an error cutoff around 3 degrees in all conditions.

an individual trial level should be similar to the correlations above). Model comparison revealed a significant interaction between condition and bot error magnitude ($\chi^2(2) = 162.49$, $p < .001$), consistent with the fact that the correlations computed above between intervention magnitude and error magnitude differed significantly by condition. Broadly, these results suggest a strong relationship between bot error and participants' intervention magnitude, particularly in the *heavy bias* and *light bias* conditions.

3.6.2 People intervened using the structure of partner error

Our results so far indicate that participants were able to correct for the potential bias exhibited by their bot partner and were even sensitive to the *continuous* variation in the magnitude of the bias. However, it remains unclear what kind of social inference about their partner, if any, participants relied on to accomplish this. If they inferred that their partner exhibited a systematic pattern of over- or under-estimating the ball's path, they could use this information to aid in correcting the bot's suggestions. However, it's also possible that participants had a roughly accurate internal model of the physics of the game which allowed them to correct the bot's error without reasoning about the underlying structure of these errors. To compare these possibilities, we examine participants' behavior on the mystery round trials; because these trials hampered participants' ability to respond using their own internal physics, success on these trials is more likely a result of inferences about the structure in their bot partner's error.

We first explore participants' intervention rate on mystery round trials. Figure 3.7 (Top Left) shows the average proportion of mystery round trials in which participants intervened on their bot partner's suggestion. While participants in the *no bias* condition rarely modified their partner's suggestion in the mystery round trials, intervention rates in the *heavy bias* and *light bias* conditions were very high. To estimate these values more precisely, we fit a generalized linear mixed effects model of participant interventions (binary) on mystery round trials with a random intercept for each subject. In this model,

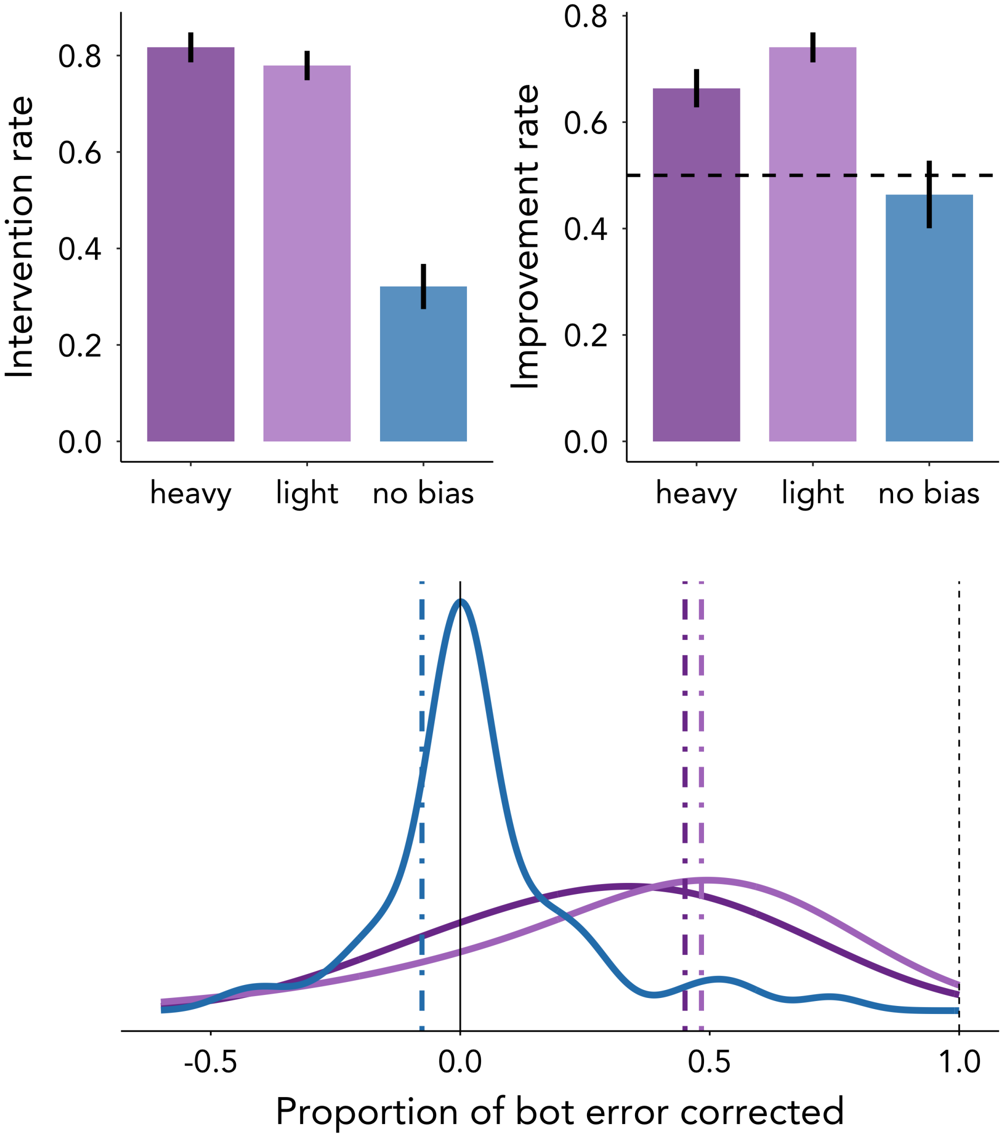


Figure 3.7. Intervention behavior on mystery round trials. (Top) At left, proportion of mystery round trials in which participants intervened to modify the bot's suggestion in each condition. At right, proportion of mystery round trials among those participants intervened in that were modified in the correct direction. The dashed line indicates chance performance (50%). Error bars represent standard error of participant proportions in both figures. (Bottom) Distribution of average error correction proportions in each condition for mystery round trials only, after removing trials in the bottom 20% of bot error. Vertical lines indicate medians.

estimated marginal mean intervention rates exhibited even more dramatic differences than the averages shown in Figure 3.7 (*heavy bias*: mean = 91.76%, 95% CI = [84.94%, 95.67%]; *light bias*: mean = 88.29%, 95% CI = [80.38%, 93.30%]; *no bias*: mean = 20.92%, 95% CI = [12.46%, 33.09%]). Intervention rate on mystery round trials differed significantly between conditions ($\chi^2(2) = 77.86$, $p < .001$); follow-up pairwise comparison of the estimated marginal means above found significant differences between the *no bias* condition and the *heavy bias* and *light bias* conditions ($p < .001$ in both cases) but no difference in mystery round intervention rate between *heavy bias* and *light bias* conditions ($p = .65$). In the absence of information about where the ball was being launched from, participants' similarly high intervention rate in the *heavy bias* and *light bias* conditions and low intervention rate in the *no bias* condition suggests that they had at least a basic representation of their bot partner's overall competence.

But did this representation extend beyond the mere need to intervene on these trials? Modifications to the bot's suggestion on a given trial can be classified as either moving the paddle in the *correct* direction (towards where the ball will land) or the *incorrect* direction (away from where the ball will land). Critically, in the *heavy bias* and *light bias* conditions, this direction on a given trial can be inferred using only knowledge of the bot's bias. Did participants understand that their partner's errors necessitated adjusting the paddle in a reliable *direction* in each half of the circle? Figure 3.7 (Top Right) plots the average proportion of mystery round trial interventions that participants modified in the *correct* direction in each condition. If participants were merely guessing, they should only get this right on roughly 50% of trials where they intervene, while bot suggestion *improvement rates* greater than 50% suggest that participants were able to recognize the directional structure of their partner's errors. To compare the improvement rates across conditions, we fit a generalized linear mixed effects model to mystery round trials in which participants intervened with improvement or worsening of the bot's suggestion (binary) as the dependent variable and a random intercept for each subject. Consistent with the average improvement rates shown in Figure 3.7, estimated marginal mean improvement rates were high in the *heavy bias* (mean = 71.24%, 95% CI = [64.66%, 77.05%]) and *light bias* conditions (mean = 77.19%, 95% CI = [71.38%, 82.11%]) and near chance in the *no bias* condition (mean = 46.16%, 95% CI = [35.78%, 56.88%]). Thus, condition was a significant predictor of whether interventions *improved* the bot's original suggestions ($\chi^2(2) = 25.19$, $p < .001$); pairwise comparison of estimated marginal means showed that this was a result of significant differences between the *no bias* and the *heavy bias* and *light bias* conditions ($p < .001$ for both), but not between the *heavy bias* and *light bias* conditions ($p = .32$). Broadly, participants not only intervened at rates that reflected their partner's ability; their interventions were more often in the right direction when the necessary intervention direction could be inferred from the bot's pattern of errors.

In light of the fact that participants in the *heavy bias* and *light bias* conditions frequently intervened in the mystery round trials, and did so in the *correct direction*, a natural question is whether the *magnitude* of their interventions was calibrated to the bot's error on these trials. We address this by once again analyzing the *proportion of bot error corrected* by participants in each condition, this time looking exclusively at mystery round trials. Figure 3.7 (Bottom) shows the distribution of participants' average proportion of bot error corrected on mystery round trials. As in the previous analysis with this measure, we remove trials in which bot error is below the the bottom 20th percentile. The distributions are qualitatively similar to those estimated for normal trials in Figure 3.6, but attenuated in the mystery round trials. Participants in the *no bias* condition once again primarily made no changes to their partner's error; changes they did make were symmetrical with respect to whether they improved the bot's suggestion. Meanwhile, participants in the *heavy bias* and *light bias* conditions showed a greater tendency to improve on their bot partner's original suggestion. We fit a linear mixed effects model to the proportion of bot error corrected for participants' mystery round trials with a random intercept for each subject, once again removing trials where bot error was below the 20th percentile of all trials in each condition. Individual bot error correction proportions on mystery round trials differed significantly across conditions ($\chi^2(2) = 39.41$, $p < .001$). As in previous analyses, estimated marginal mean pairwise comparisons revealed that this was a result of significant differences between the *no bias* and the *heavy bias* and *light bias* conditions ($p < .001$ in both cases) but not between the *heavy bias* and *light bias* conditions ($p = .68$). Estimated marginal means for the proportion of bot error corrected reflect these differences: Mean proportion corrected was 29.95% for the *heavy bias* condition (95% CI = [20.30%, 39.55%]), 35.53% for the *light bias* condition (95% CI = [26.50%, 44.55%]), and -5.39% for the *no bias* condition (95% CI = [-14.90%, 4.09%]). Thus, participants in the bias conditions not only modified their bot partner's suggestions in the right *direction*, but did so in a way that approximated the *magnitude* of their partner's error.

3.7 Discussion

In this experiment, we explored people's ability to collaborate with a partner whose internal model of the task itself was fundamentally different from the ground truth, leading to systematic patterns of error. Results contain two key findings. First, we show that participants successfully collaborated with *biased* bots—they intervened frequently with biased partners and achieved error levels similar to those in a control condition with a highly accurate, unbiased bot. Furthermore, we find evidence that participants' ability to correct their partner's bias was sensitive to the *continuous*, graded nature of the bias; participants primarily corrected their partner's errors in the right direction and there was a strong relationship between magnitude of the partner's error and the magnitude of participants' corresponding interventions.

Our second finding concerns the underlying *competence evaluations* that supported participants' adaptive behavior. One account of participants' success on the task is that they were merely adjusting the paddle based on where they expected the ball to land, without any reasoning about their partner. To explore whether participants could infer the structure underlying their partner's suggestions and use this to guide behavior, participants were presented with eight *mystery round* trials in which the ball's launch point was hidden from them. Presented with the agent's (informed) paddle location suggestion but not the ball's launch point, participants needed to decide where to place the paddle based on what they knew about the relationship between the agent's suggestions and the ball's landing point. We find that participants paired with both biased bots (but not the unbiased bot) were significantly more likely to intervene on these trials and, when doing so, adjusted the paddle in the correct direction. This ability to move the paddle in the direction which corrected for the bot's bias without knowledge about where the ball was being launched from in the first place is difficult to explain without some understanding of the underlying structure in the bot's suggestions.

The current experiment therefore provides evidence that people's competence evaluations—their understanding of what their bot partner is capable of—extend beyond mere assessments of accuracy or variance to recognizing structure in their errors that result from an inaccurate model of the task. The ability to detect this structure and use it to inform their collaboration with their partner places the current results closer to many everyday evaluations of competence, in which we seek to develop a more comprehensive understanding of other people's *internal model* of an activity.

3.8 General Discussion

What kind of representation do people form of others' abilities, and how do these representations impact their collaborative behavior? Prior work has addressed this question from a range of distinct theoretical and methodological perspectives across psychology and artificial intelligence; broadly, this work emphasizes people's ability to incorporate sophisticated representations of the task as well as their collaborator's knowledge, prior success, and strategy. The current work builds on these results in two key ways. First, while prior research has largely relied on collaborators whose abilities are static, this contrasts with many everyday interactions in which people can improve with practice or worsen with fatigue. Intuitively, our representations of others' competence are flexible enough to accommodate this, yet the ways in which people represent an *improving* or *worsening* partner remain largely unclear. Second, while some prior work has explored the impact of people's underlying task representations on judgments of competence or trust (Soh et al., 2020), existing work tends to focus on collaborative interactions in fairly abstract domains. In the current work, participants' collaborative interactions occurred in the context of a physics-based game in which people were likely to have rich mental models of the basic mechanics of gravity and ball trajectories. In experiment 2, we leveraged this to modify the bot agent's internal model of the task, a manipulation which would be challenging in more abstract domains. Further, whereas people's own sparse internal models may make it harder to detect this sort of manipulation in abstract tasks, we hypothesized that people's flexible ability to reason about physical interactions would enable them to more easily detect their partner's systematic error patterns.

Our results show that both of these key manipulations captured people's intuitive reasoning about the abilities of others in collaborative contexts. First, in experiment 1, we find that people adapted their intervention behavior to accommodate the underlying accuracy—and the *changes in accuracy*—of their partner. Participants were sensitive to changes in the variance of their partner's suggestions over time. Further, results from the *critical trials* in experiment 1 indicate that people's collaborative behavior was supported by a *latent estimate* of their partner's ability which influenced their interventions above and beyond trial-level information. Put another way, we show that people incorporated the past behavior of their partner into their intervention decisions. Future work should explore this relationship more closely; how much does history matter?

Second, in experiment 2, we find that people were sensitive to their partner's incorrect model of the task mechanics. Specifically, across the two bias conditions and the *no bias* control condition, people were able to achieve high accuracy levels; in particular, participants paired with the biased bots showed evidence of systematically correcting for the varying magnitude of their partner's errors. Critically, our results also suggest that correcting their partner's errors relied, at least in part, on the ability to represent their partner's incorrect world model. Participants intervened frequently, and overwhelmingly in the correct direction, on trials in which the only information about the ball's launch point (other than the quadrant of the circle) came from the bot's suggestion itself.

Taken together, these results suggest that physical task settings offer a promising domain in which to explore the representations of competence that underlie our collaborative behavior. However, the current work is only a first step in this direction. People exhibit a rich ability to simulate physical environments in service of prediction and explanation (Battaglia et al., 2013) and rely on similarly rich mental models of others' motivation and actions (Jara-Ettinger et al., 2016). The current work presents an opportunity to explore how these two systems might interact when forming representations of others' knowledge or abilities in physical task settings. To what degree can we simulate the physical knowledge of others using our own internal physics? Or recognize an agent's goals based on their interaction with the physical world?

Future work might also expand on the kinds of latent parameters that people use to represent others' abilities alongside more detailed representations of competence. For example, can people's *trust* in another agent be represented by a single parameter estimated from their previous behavior, similar to accuracy? Or is it more like a summary statistic over the full mental model we have of others, describing them on the basis of everything that is known about their knowledge and abilities in a domain? Though existing work offers few answers to such questions, the ability to probe people's behavior over many interactions in rich task domains, and recent advances in computational modeling of social inferences (Vélez & Gweon, 2021) offers a promising avenue for investigation.

Finally, though intuition and prior work on pedagogical reasoning (Shafto et al., 2014) suggests that people are highly attuned to the learning goals of others, the current work is only a first step in understanding our representations of others' learning processes. Physical tasks which can be decomposed into relevant physical knowledge and competencies (like knowing the correct mass of the ball in the current task) offer a chance to better explore how people form representations of another agent's learning in more discrete and structured ways over the course of many repeated interactions. Altogether, current results open the door to future work exploring how our representations of others' abilities, knowledge, and learning enable rich forms of collaboration over extended time periods.

3.9 Acknowledgments

I would like to thank my co-authors for their feedback and invaluable contributions to this manuscript: Haoliang Wang, Justin Yang, Suvir Mirchandani, Erdem Bıyık, Dorsa Sadigh, and the inimitable Judith Fan.

Chapter 3, in full, is currently being prepared for submission for publication of the material. An earlier version of the project was published as Brockbank, E., Wang, H., Yang, J., Mirchandani, S., Bıyık, E., Sadigh, D., & Fan, J. E.. (2022). How do people incorporate advice from artificial agents when making physical judgments? In J. Culbertson, A. Perfors, H. Rabagliati, & V. Ramenzoni (Eds.), *Proceedings of the 44th Annual Meeting of the Cognitive Science Society* (pp. 1469-1475). Cognitive Science Society. The dissertation author was the primary investigator and author of this material.

References

Aboody, R., Velez-Ginorio, J., Santos, L. R., & Jara-Ettinger, J. (2023). When naïve pedagogy breaks down: Adults rationally decide how to teach, but misrepresent learners' beliefs. *Cognitive Science*, *47*(3), e13257.

Bahrami, B., Olsen, K., Latham, P. E., Roepstorff, A., Rees, G., & Frith, C. D. (2010). Optimally interacting minds. *Science*, *329*(5995), 1081–1085.

Bass, I., Bonawitz, E., Hawthorne-Madell, D., Vong, W. K., Goodman, N. D., & Gweon, H. (2022). The effects of information utility and teachers' knowledge on evaluations of under-informative pedagogy across development. *Cognition*, *222*, 104999.

Bass, I., Gopnik, A., Hanson, M., Ramarajan, D., Shafto, P., Wellman, H., & Bonawitz, E. (2019). Children's developing theory of mind and pedagogical evidence selection. *Developmental psychology*, *55*(2), 286.

Battaglia, P. W., Hamrick, J. B., & Tenenbaum, J. B. (2013). Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, *110*(45), 18327–18332.

Biele, G., Rieskamp, J., & Gonzalez, R. (2009). Computational models for the combination of advice and individual learning. *Cognitive science*, *33*(2), 206–242.

Bonawitz, E., Shafto, P., Gweon, H., Goodman, N. D., Spelke, E., & Schulz, L. (2011). The double-edged sword of pedagogy: Instruction limits spontaneous exploration and discovery. *Cognition*, *120*(3), 322–330.

Chen, M., Nikolaidis, S., Soh, H., Hsu, D., & Srinivasa, S. (2020). Trust-aware decision making for human-robot collaboration: Model learning and planning. *ACM Transactions on Human-Robot Interaction (THRI)*, *9*(2), 1–23.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences.* Academic press.

Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, *415*(6870), 429–433.

Gweon, H. (2021). Inferential social learning: Cognitive foundations of human social learning and teaching. *Trends in Cognitive Sciences*, *25*(10), 896–910.

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.

Jara-Ettinger, J., Tenenbaum, J. B., & Schulz, L. E. (2015). Not so innocent: Toddlers' inferences about costs and culpability. *Psychological science*, *26*(5), 633–640.

Leonard, J. A., Bennett-Pierre, G., & Gweon, H. (2019). Who is better? preschoolers infer relative competence based on efficiency of process and quality of outcome. *CogSci*, 639–645.

Leong, Y. C., & Zaki, J. (2018). Unrealistic optimism in advice taking: A computational account. *Journal of Experimental Psychology: General*, *147*(2), 170.

Pescetelli, N., & Yeung, N. (2021). The role of decision confidence in advice-taking and trust formation. *Journal of Experimental Psychology: General*, *150*(3), 507.

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.

Soh, H., Xie, Y., Chen, M., & Hsu, D. (2020). Multi-task trust transfer for human-robot interaction. *The International Journal of Robotics Research*, *39*(2-3), 233–249.

Staudinger, M. R., & Büchel, C. (2013). How initial confirmatory experience potentiates the detrimental influence of bad advice. *NeuroImage*, *76*, 125–133.

Vélez, N., & Gweon, H. (2019). Integrating incomplete information with imperfect advice. *Topics in cognitive science*, *11*(2), 299–315.

Vélez, N., & Gweon, H. (2021). Learning from other minds: An optimistic critique of reinforcement learning models of social learning. *Current Opinion in Behavioral Sciences*, *38*, 110–115.

Xie, Y., Bodala, I. P., Ong, D. C., Hsu, D., & Soh, H. (2019). Robot capability and intention in trust-based decisions across tasks. *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 39–47.

Chapter 4

Discussion

The work in this dissertation is motivated by the broad question of *how our representations of others are structured*. While answering this question is beyond the scope of one dissertation (or many!), I argued at the outset that we can get traction on the question—and on areas where current research leaves us guessing—by turning to the history of how psychologists have thought about good scientific models of others. In particular, I argued that *behaviorist* and *cognitivist* models of human reasoning offer two well-described accounts of how we might think about representations of others in *intuitive psychology*. Further, I argued that this distinction offers a useful way to evaluate existing work on theory of mind reasoning and social reasoning more broadly. A review of the literature on our *behaviorist* and *cognitivist* intuitive psychology suggests that while recent developments in computational cognitive modeling have enabled a systematic inquiry into the structure of our cognitivist representations of others as well as their ontogeny, the space of behaviorist mental models is disparate and lacks a coherent account. While a range of everyday human behaviors, from habits to sequenced actions to norms and conventions might seem like candidates for a more "rule-governed" theory of mind, little work has systematically explored people's ability to reason about others in this way.

The first two chapters of the dissertation represent my own attempt to characterize how people reason about structured sequential patterns in others' behavior. In chapter 1, I show that people's behavior in *mixed strategy equilibrium* (MSE) games like rock, paper, scissors represent a novel setting in which to do this for several reasons. First, the structure of the game itself entails that looking for sequential patterns in an opponent's behavior is the only strategy available, assuming a fallible opponent. Second, the simple action space of rock, paper, scissors allows for a systematic account of the different sequential patterns that a player can display and that their opponent might in turn exploit. And third, people can play many repeated rounds of the game, allowing ample time for players to express such patterns and detect them in others. I present evidence that people are able to exploit their opponents in this way over many rounds and illustrate the patterns that emerge in their moves when playing.

Chapter 2 builds on these results by attempting to characterize precisely what it is people are doing when they acquire a *behaviorist opponent model* in this setting. Over two studies in which people played repeated rounds against bot opponents that either exhibited particular patterns in their moves or tried to exploit those same patterns in participants' moves, I show that there is a highly constrained space of patterns people can reliably exploit or revise in their own moves. Thus, people's behaviorist mental models appear to have a limited capacity in this setting. This is counterintuitive and goes against prior work showing that people's statistical learning abilities are rich and sophisticated (Saffran et al., 1996); it merits future work which might systematically tease apart the reasons people struggle to reason about others' behavior in this task.

However, one approach is by exploring what makes people *succeed* at developing more cognitivist models of intuitive psychology. In chapter 3, I turn to the question of how people develop representations of another agent's *competence* to support repeated collaboration. This question has relevance to existing work spanning developmental psychology, social psychology, and robotics and artificial intelligence. Recent computational work exploring our social learning suggests that people can build rich and highly structured mental models of other agents, particularly in service of collaborative goals (Gweon, 2021; Vélez & Gweon, 2021). Chapter 3 extends earlier work in this vein by exploring people's ability to collaborate with agents whose abilities change over time and who exhibit incorrect internal models of the task. Despite the difference in topics, the methodological approach in this chapter, and the motivating interest in understanding the precise structure of our representations of others as they unfold over time, is closely aligned with chapters 1 and 2. Using repeated interactions with a bot partner whose behavior was parametrically manipulated in broad ways, my co-authors and I show that people develop a latent representation of their partner's ability based on past trials which influences their collaborative behavior and that this representation can be extended to include features of their partner’s own internal task model.

What do we learn from this work collectively? In one sense, the results from chapters 1 and 2 stand in fairly stark contrast to those in chapter 3. While people struggled to develop complex behaviorist mental models of their rock, paper, scissors opponents, they made nuanced and sophisticated inferences about the structure in their partner’s behavior in a collaborative physical task. In light of this difference, one apparent take-away from these results is that **our representations of others can take on surprisingly diverse forms but their complexity is heavily context-dependent**. When collaborating in a physical task domain, participants built rich models of others’ behavior; in adversarial interactions with an abstract action space, people struggled to infer the causes of their opponent’s behavior beyond simple contingencies. There are a number of possible accounts for these differences that could be explored in future work: the difference between collaboration and competition, the difference between abstract, game-like action spaces and more grounded, physical actions, the type of pattern being exhibited (sequential patterns versus bias and variance), or the error signal acquired from actions (win, loss, or tie versus continuous error).

However, an important take-away from this work is not just the differences between results, but the fact that people developed well-defined mental models of others in all of these settings. This was by no means a guarantee and may be surprising given the large set of differences outlined above. Thus, **a central question that arises from these results is how people are able to determine the right kind of mental model of others for a given context**. The idea that people deploy mental models of others at different levels of granularity or complexity has been explored in recent work (Burger & Jara-Ettinger, 2020; Rabinowitz et al., 2018), but what constitutes the hypothesis space or the proper parameterization of these distinct models is far from clear. The work in this dissertation offers one possible way forward: By identifying domains in which people can and do exhibit distinct mental models of others and then experimentally manipulating the differences between these domains—the costs, time constraints, and complexity of others' behavior—we may be able to develop a more holistic account of how people weigh the tradeoffs between computational complexity and predictive power when constructing representations of others.

A second question that emerges from this work is how context-specific people's representations of others are, along with the learning and inference mechanisms that support them. One possible account of people's failures in detecting patterned opponent moves in rock, paper, scissors, is that people struggle to recognize patterns in abstract domains that they might easily identify in more familiar settings. For example, Cheng and Holyoak (1985) famously showed that when a difficult propositional logic task was reframed as a question about who should be asked to show their ID at a bar, people found it trivially easy. They argue that *permission schemas* offer a means of reasoning that the logical formulation of the original problem did not permit. Do the schemas we invoke for interpreting others' actions show a similar selectivity where sequential patterns in behavior that evoke *preferences* or other familiar abstractions are easily recognizable, while the same patterns expressed in rock, paper, scissors moves seem inscrutable? In fact, questions about context-specificity arise in chapter 3 as well. One intuitive aspect of our everyday inferences about the competence of others is that they do not happen in a vacuum; if a child is able to brush their teeth on their own, they can probably get dressed on their own as well. How do we perform such mappings between everyday contexts, and where do they break down? What is the structure of our everyday *task embeddings* and how do they support our representations of others' abilities within particular contexts?

Finally, a third question that arises from this work is how we might further characterize the behaviorist and cognitivist mental models explored in these chapters. First, consider the behaviorist intuitive psychology outlined in chapters 1 and 2. While sequential patterns in an opponent's RPS moves represent a simple and structured way to explore such reasoning, everyday life is full of scenarios in which people follow well-defined *rules* or *scripts* for behavior (Schank & Abelson, 1977). Consider, for example, the actions of a barista taking somebody's order in a coffee shop. Might the notion of a *behaviorist* theory of mind be expanded to include the more general case of how we reason about others' behavior as a result of such well-defined rules or procedures? How much of our everyday reasoning about others falls under this category and how distinct is it from existing cognitivist mental models?

Second, when thinking about the representation that participants formed of their bot partner in chapter 3 (experiment 1), it likely centered on a simple trait estimate like how *good* or *accurate* the agent was. Recent computational accounts of theory of mind such as the naïve utility calculus model focus on how we infer relatively *transient* mental states like beliefs and desires, but largely ignore more stable and individually varying *traits* like whether somebody is impatient or careless (Jara-Ettinger, 2019). Yet traits play an important role in our intuitive psychology—when somebody cuts us off on the highway, we accuse them of being negligent rather than assume that they are in a hurry—and the relationship between traits and more context-specific causes has been one of the most well-studied phenomena in social psychology (Jones & Harris, 1967; Ross, 1977; Walker et al., 2015). Recent work has explored how we learn useful trait-like abstractions from repeated interactions with others (Hackel & Amodio, 2018; Tamir & Thornton, 2018; van Baar et al., 2022); how might this be extended to capture the range of trait inferences people routinely make about those around them and how do these inferences support predictions of future behavior?

I look forward to the opportunity to explore these questions and other similar ones in future work (hopefully with members of this committee!), and thank all of you for your time, support, and feedback along the way.

References

Burger, L., & Jara-Ettinger, J. (2020). Mental inference: Mind perception as bayesian model selection. *CogSci*.

Cheng, P. W., & Holyoak, K. J. (1985). Pragmatic reasoning schemas. *Cognitive psychology*, *17*(4), 391–416.

Gweon, H. (2021). Inferential social learning: Cognitive foundations of human social learning and teaching. *Trends in Cognitive Sciences*, *25*(10), 896–910.

Hackel, L. M., & Amodio, D. M. (2018). Computational neuroscience approaches to social cognition. *Current Opinion in Psychology*, *24*, 92–97.

Jara-Ettinger, J. (2019). Theory of mind as inverse reinforcement learning. *Current Opinion in Behavioral Sciences*, *29*, 105–110.

Jones, E. E., & Harris, V. A. (1967). The attribution of attitudes. *Journal of experimental social psychology*, *3*(1), 1–24.

Rabinowitz, N., Perbet, F., Song, F., Zhang, C., Eslami, S. A., & Botvinick, M. (2018). Machine theory of mind. *International conference on machine learning*, 4218–4227.

Ross, L. (1977). The intuitive psychologist and his shortcomings: Distortions in the attribution process. In *Advances in experimental social psychology* (pp. 173–220). Elsevier.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926–1928.

Schank, R. C., & Abelson, R. P. (1977). *Scripts, plans, goals, and understanding: An inquiry into human knowledge structures*. Routledge.

Tamir, D. I., & Thornton, M. A. (2018). Modeling the predictive social mind. *Trends in cognitive sciences*, *22*(3), 201–212.

van Baar, J. M., Nassar, M. R., Deng, W., & FeldmanHall, O. (2022). Latent motives guide structure learning during adaptive social choice. *Nature Human Behaviour*, *6*(3), 404–414.

Vélez, N., & Gweon, H. (2021). Learning from other minds: An optimistic critique of reinforcement learning models of social learning. *Current Opinion in Behavioral Sciences*, *38*, 110–115.

Walker, D., Smith, K. A., & Vul, E. (2015). The'fundamental attribution error'is rational in an uncertain world. *CogSci*.