Integrating Human Patterning Behaviors for the Improvement of Human-Robot Teams

by

Clare M. Lohrmann

B.S., George Washington University, 2018

M.S., University of North Carolina Wilmington, 2019

A thesis submitted to the

Faculty of the Graduate School of the

University of Colorado in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

Department of Computer Science

2026

Committee Members:

Bradley Hayes, Chair

Alessandro Roncone

Christoffer Heckman

Nisar Ahmed

Chien-Ming Huang

Lohrmann, Clare M. (Ph.D., Computer Science)

Integrating Human Patterning Behaviors for the Improvement of Human-Robot Teams

Thesis directed by Prof. Bradley Hayes

As human-robot teams become more prevalent in domains such as disaster response, autonomous transportation, and large-scale logistics, a central challenge is enabling humans to effectively interpret and predict the behavior of robotic agents. While advances in autonomy have improved robotic coordination, human teammates often remain limited by cognitive constraints when attempting to understand complex, dynamic systems. This creates a critical bottleneck in collaboration: as the number of agents increases or environments become more complex, human oversight and decision-making become increasingly difficult. Addressing this challenge requires approaches that support human understanding without sacrificing the flexibility or autonomy of robotic algorithms.

Existing work in human-robot interaction has largely focused on improving transparency through explanations, visualizations, or communication interfaces. However, these approaches often emphasize increased communication rather than the collective structure of the system. What remains underexplored is how the behavior of a robotic system itself can be intentionally shaped to make it more interpretable to human observers. Specifically, little is known about whether introducing higher-order behavioral patterns into human-robot teams can reduce perceived complexity and improve human predictive performance.

This thesis investigates whether patterning can enhance human understanding in human-robot teams by restructuring how complex group behavior is perceived and processed.

Dedication

To my family, whose unwavering support, patience, and encouragement made this journey possible.

To my peers, for the camaraderie, collaboration, and shared perseverance that transformed challenges into growth.

To my advisors, whose guidance, expertise, and belief in my work shaped both this thesis and my development as a researcher.

And to the ducks; an unexpected but constant source of perspective, levity, and quiet companionship throughout the process.

This work is dedicated to all of you.

Acknowledgements

I would like to express my sincere gratitude to everyone who contributed to the completion of this thesis and supported me throughout this journey.

First and foremost, I would like to thank my advisors, Bradley Hayes and Alessandro Roncone, for their mentorship, expertise, and steadfast support throughout this process. Their insight, patience, and encouragement challenged me to think critically and grow as a researcher. I am deeply grateful for the time and care they invested in both this work and my development.

I would also like to extend my appreciation to Christoffer Heckman, Nisar Ahmed, and Chien-Ming Huang for their valuable feedback, thoughtful perspectives, and contributions to this research. Their guidance strengthened this thesis and broadened my understanding of the field.

To my peers and colleagues, including Maria, Breanne, Emily, and Christine, thank you for the collaboration, discussions, and camaraderie that made this experience both productive and meaningful. The shared challenges, ideas, and support created a community that I will always value. I am grateful to CU for providing the resources, environment, and opportunities that made this work possible. I would also like to acknowledge the Army Research Lab, whose support contributed to the successful completion of this research.

To my family, thank you for your unwavering encouragement, patience, and belief in me. Your support sustained me through every stage of this journey, and this achievement is as much yours as it is mine.

This thesis represents not only my own efforts, but the collective support of all those who helped me along the way. For that, I am deeply thankful.

Contents

Chapter

Tables

Table

Figures

Figure

Chapter 1

Introduction

Robots are becoming increasingly capable of operating alongside humans in complex, dynamic environments, creating new opportunities for collaborative human–robot systems. As these systems grow in scale and autonomy, designing interactions that humans can readily interpret and coordinate with becomes an increasingly important challenge. This dissertation explores how structured behavioral patterns can improve predictability and reduce effective complexity in human–robot teams, enabling more scalable and intuitive coordination.

1.0.1 Motivation

Human-AI teaming is fundamentally shaped by a persistent tension between machine optimization and human cognition. Autonomous systems can generate highly efficient behaviors, but these behaviors are often difficult for people to understand, anticipate, and coordinate with because they are not grounded in the same cognitive mechanisms humans use to interpret the behavior of other humans. People rely on heuristics, patterns, and mental models to make sense of behavior, rather than solving complex optimization problems[45, 85, 1]. As a result, even highly capable autonomous systems can appear unpredictable to human partners, creating a "predictability gap" that degrades trust, coordination, and overall team performance. This challenge is not unique to robotics, but reflects a broader issue in human-autonomy interaction: across domains, machine-optimal actions can conflict with human expectations and make collaboration more difficult.

In human-robot interaction, this problem is especially acute because effective teaming de-

pends on a human's ability to form an accurate mental model of their robotic partner. A mental model, a cognitive structure that organizes knowledge about how a system or person functions, enables a person to infer the robot's intentions, anticipate its future actions, and adapt their own behavior accordingly[56, 125, 122]. Prior work has consistently shown that predictable robot behavior improves trust, fluency, and team effectiveness, while unpredictable behavior can hinder performance even in simple tasks[27, 36]. Although substantial research has focused on improving a robot's ability to predict and adapt to human actions, less attention has been given to the inverse problem: enabling humans to better predict robots. Yet mutual predictability is essential for collaboration, and this remains a critical technical and cognitive gap.

This gap can be understood by comparing human-robot teams to human-human teams. Humans routinely coordinate effectively in complex collaborative tasks because they share common cognitive tools for interpreting behavior, including heuristics, pattern recognition, and sensitivity to repeated structure[45, 40, 3, 78, 85, 1, 68, 80]. These mechanisms allow people to quickly identify intent and anticipate future actions, and help to form an individual's mental model[56, 80, 40, 3]. By contrast, robots usually generate behavior through optimization-based methods that fundamentally differ from how humans reason. Reinforcement learning and related approaches are highly effective at maximizing task reward, but they typically prioritize locally optimal actions without explicitly considering how the resulting sequence of behavior will be perceived by a human observer. Because these methods are often rooted in Markovian assumptions and immediate reward maximization, they can produce behaviors that are efficient but difficult to interpret as a whole. Additionally, information asymmetry, which is inherent to human-robot interaction, complicates this further, as humans and robots are reasoning over different data.

This thesis argues that improving human-robot teaming requires a shift in design philosophy: predictability must be treated as a primary design objective rather than an incidental property of efficient behavior. Specifically, this work is motivated by the hypothesis that embedding human-perceptible structure into robot behavior can significantly improve a human teammate's ability to understand and anticipate the robot's actions, even if doing so requires suboptimality on the

part of the robot. Rather than viewing patterns as constraints that reduce efficiency, this thesis frames them as tools for cognitive alignment within the context of human-robot teaming. The core premise is that autonomous systems should not only act effectively, but should behave in ways that are predictable to the people working alongside them.

To support this argument, this thesis explores how patterns can be explicitly encoded into robot decision-making. Humans are deeply adept at recognizing and using patterns across domains, from visual sequences to temporal rhythms, and this capacity offers a powerful opportunity for improving robot predictability[95]. By structuring robot behavior around patterns created from human-perceptible features such as object properties, task order, or spatial structure, it becomes possible to generate behaviors that are easier for humans to model and anticipate. This thesis introduces methods for identifying and selecting patterns at the task and navigational levels, as well as on a group level, allowing robots to preserve strong task performance while improving transparency and coordination.

More broadly, this work advances a framework for reconciling the strengths of machine optimization with the realities of human cognition. By integrating human-centered structure into planning and control, this thesis contributes toward a broader vision of autonomous systems that are understandable, trustworthy, and effective partners in shared environments by centering human cognition and reasoning.

1.0.2 Research Questions

In collaborative settings, humans rely extensively on shared knowledge and models to coordinate behavior. A key mechanism that simplifies collaboration is the presence of patterns: predictable sequences of actions that people can identify, learn, and generalize across contexts [95]. Humans are deeply adept at recognizing patterns, and this ability plays a central role in how they interpret behavior, infer intent, and form expectations about what others will do next [80]. In human-human teams, this shared reliance on patterns and conventions supports fluent coordination even in dynamic and uncertain environments[56, 125, 122].

A central premise of this dissertation is that these same cognitive mechanisms can be leveraged to improve human-robot collaboration. If humans naturally rely on patterns to understand and predict others, then robot behavior that is structured around perceivable patterns may be easier for people to interpret and anticipate.

This motivation gives rise to three core research questions. First, how can patterns be formally represented in the context of robotic planning? While humans intuitively recognize patterns, robotics lacks a principled framework for defining, measuring, and reasoning about the kinds of structure that are meaningful to people. Addressing this question requires translating the abstract concept of a pattern into a computational representation that can be embedded within planning and control systems.

Second, once patterns are formalized, how can robot behavior be adjusted to follow them? Existing planning and control methods are typically designed to optimize efficiency, safety, or reward, often without regard for how the resulting behavior is perceived by a human observer. Yet effective teaming requires robots to reason not only about task success, but also about the structure of their actions over time. This dissertation therefore examines how patterns can be incorporated into robotic decision-making in a principled way, enabling robots to produce behavior that remains effective while becoming more consistent, interpretable, and predictable.

Third, how does patterned robot behavior affect human ability to predict robots and team performance? The central hypothesis explored throughout this dissertation is that embedding perceivable structure into robot actions can improve a human teammate's ability to form accurate mental models, anticipate future actions, and coordinate more effectively. This raises broader questions about the tradeoff between predictability and task performance: how much structure is needed to meaningfully improve teaming, what kinds of patterns are most useful, and whether modest sacrifices in individual robot efficiency may produce disproportionately large gains in team fluency.

Taken together, this dissertation investigates how the human cognitive strength of pattern recognition and abstraction can inform the design of autonomous systems. By examining how

patterns can be formalized, embedded into robot behavior, and evaluated in the context of human teaming, this work aims to advance a broader vision of autonomy that is not only capable and efficient, but also understandable, predictable, and well aligned with the people it is designed to support.

1.0.3 Contributions

This thesis investigates how human-perceptible patterns can be used to improve coordination in human-autonomy teams. The first two works make foundational contributions by formalizing patterns as a mechanism for shaping autonomous behavior in two complementary settings: discrete task planning and continuous control. In the discrete domain, we introduce Pattern-Aware Convention-setting for Teaming (PACT), a filtering algorithm that enables embodied agents to adopt human-perceptible behavioral patterns that improve predictability during collaboration. In the continuous domain, we develop PRESTO, a framework that formalizes patterned behavior in continuous action spaces and explicitly characterizes the tradeoff between predictability and optimality. Together, these works establish both the algorithmic foundations and empirical evidence needed to understand how structured, human-legible behavior can improve team performance. Building on this foundation, the final component of this thesis extends these ideas to multi-agent systems, exploring how social patterns such as dynamic grouping can improve predictability, reduce cognitive load, and increase the number of autonomous agents a human can effectively manage.

Our results demonstrate that incorporating pattern structure into robot behavior yields substantial benefits. Using PACT, robots become more predictable to human teammates, leading to improvements in both objective team performance and subjective perceptions of the robot. These findings support the broader hypothesis that aligning robot behavior with human cognitive strengths enables people to more readily infer intent, understand decision-making processes, and generalize behavior across contexts. More broadly, this work highlights the value of designing autonomous systems that are not only effective, but also cognitively compatible with their human partners.

To more deeply understand the relationship between predictability and performance, we leverage PRESTO as a scientific instrument to explicitly characterize this tradeoff. Our experiments reveal a striking result: modest, bounded sacrifices in optimality can produce disproportionately large gains in team fluency. We provide the first direct empirical measurement of this effect in the context of reinforcement learning policies, demonstrating that predictability plays a central role in effective human-autonomy interaction.

Across two human-subject studies, we uncover several key principles. First, we observe a decoupling between perceived intelligence and behavioral complexity, challenging the common assumption that more complex or mathematically optimal behavior is viewed as more intelligent [118, 113, 7, 53, 86]. Second, by tracking participant predictions over time, we provide quantitative evidence that patterned behavior supports the formation of more stable and accurate mental models of an autonomous partner—an essential component of effective team cognition. Together, these findings suggest that predictability, rather than raw optimality, is the primary driver of fluent human-autonomy teaming.

Finally, we extend this perspective to more complex, multi-agent settings. As the number of agents increases, so too does the cognitive burden placed on human collaborators. Incorporating pedestrian-inspired patterns does improve the ability of participants to predict agent paths, but only under higher-complexity circumstances, when the grouping patterns significantly reduce the number of agents people are building mental models of, as groups can be modeled as a singular agent.

Taken together, this work establishes a unifying perspective: patterning is a powerful mechanism for shaping robot behavior in ways that align with human cognition. PACT demonstrates how patterns can be introduced in discrete task spaces, while PRESTO extends this framework to continuous domains and formalizes the balance between predictability and optimality. Building on these foundations, we explore how social patterns can be leveraged in multi-agent systems to further enhance human-autonomy teaming and reduce cognitive load.

1.0.4 Roadmap

This dissertation is organized as follows. Chapter 1 introduces the central research problem of improving coordination and predictability in human-robot teaming, and outlines the key questions, motivations, and contributions that guide the dissertation. It also provides an overview of the broader challenges in designing robotic systems that are both effective and legible to human collaborators.

Chapter 2 presents the necessary background and related work. This chapter reviews prior research in human-robot interaction, shared autonomy, convention formation, legibility and predictability of robot behavior, and multi-agent coordination. It also establishes the theoretical foundations and technical tools that motivate the approaches developed in this dissertation.

Chapters 3 through 5 each present a core research contribution in the form of a published or submitted paper. Chapter 3 introduces the first work, which focuses on formalizing patterns at the subtask-planning level as well as validating the use of patterns in human-robot teaming, including the problem formulation, proposed method, and experimental evaluation. Chapter 4 builds on this foundation by addressing the formalization of patterns in the continuous space as well as the balance of patterning predictability and optimality, extending the framework to navigation. Chapter 5 presents the final work, which investigates the use of pedestrian patterns in multi-agent supervision.

Finally, Chapter 6 concludes the dissertation by summarizing the key findings across all three projects, discussing their collective implications for human-robot teaming, and outlining promising directions for future work. Together, these chapters demonstrate a progression from foundational concepts to increasingly complex and realistic collaborative settings, advancing the design of robotic systems that better support human partners.

Chapter 2

Background and Related Work

As human–robot teams grow in scale and autonomy, people must increasingly interpret and predict the behavior of agents in a variety of contexts. Prior work in human–robot interaction and cognitive systems has shown that effective coordination depends on the human's ability to form accurate mental models of collective behavior, yet these models are less accurate with robots and become increasingly strained as environmental and system complexity grow.

2.0.1 Human-Robot Teaming

Human–robot teaming as a subfield focuses on how humans and robots coordinate as interdependent partners to achieve shared goals. Rather than treating robots as tools or isolated agents, this line of work emphasizes joint action, mutual adaptation, and the development of shared mental models[108]. Effective teaming depends on the robot's ability to communicate intent, respond to human behavior, and maintain transparency in its actions, as well as the human's ability to anticipate and appropriately rely on the robot[116, 50]. Prior research has explored factors such as trust, fluency, and coordination efficiency, showing that teams perform best when interaction feels predictable and aligned, with both partners contributing in complementary ways[34, 116, 77]. This has motivated approaches that enable robots to exhibit legible behavior, communicate internal state, and adjust autonomy dynamically in response to human needs[31, 82, 28].

While much of the human–robot teaming literature has focused on one-to-one interactions, extending these principles to multi-agent settings introduces additional complexity. In teams in-

volving multiple robots, humans must reason not only about individual agents but also about how those agents coordinate with each other as a collective[32, 8, 12]. This can strain the formation of shared mental models and complicate trust calibration, as system behavior becomes more distributed and less transparent[12]. As a result, there remains a gap in understanding how to design robot teammates that are not only functionally effective but also cognitively aligned with human partners, particularly in terms of enabling scalable understanding and prediction as system complexity grows.

Much of the recent work in human-robot collaboration focuses exclusively on improving the performance of the robotic agent or its mental model of humans. Works that attempt to predict human actions or their path directly have seen success within the environments they tested in [98, 39, 42]. There has also been a significant effort to adapt successful methods in competitive environments to collaborative environments [58, 17, 111, 52], though this is very difficult. Approaches that are highly effective in competitive environments are challenging to adapt to collaborative environments [51]. What makes many self-play approaches successful—a policy that is convoluted and difficult for opponents to counter—is a drawback in collaborative settings. What results is a large drop in performance when trained agents are tested with humans rather than other agents [106, 58]. These approaches also do not capture the full scope of human collaboration within their environments [51].

2.0.2 Predictability and Legibility

In this thesis, we define predictability as "the quality of matching expectations", which is directly taken from foundational predictability and legibility work [31]. Going forward, it is critical to define and differentiate predictability and legibility. Predictability and legibility describe two complementary aspects of how humans interpret robot behavior, and the distinction is framed around the direction of inference between goals and actions. Legibility refers to how easily a human can infer the robot's goal or intent from observing its actions. By contrast, predictability refers to how easily a human can anticipate a robot's future actions once the robot's goal is known.

As mental models encode human expectations, creating robot behaviors that are more easily modeled by humans (creating autonomous systems that act predictably) is a prerequisite for effective human-robot collaboration [65, 31, 27, 71, 36]. Use of conventions (consistently performing an action in a given context) has also been leveraged to facilitate improved human-robot collaboration [92, 2, 41, 73, 20], and similarly relies on humans recognizing repeated robot behavior. Empirical studies within the HRI literature have repeatedly validated the premise that humans find predictable robots easier and more satisfying to work with; humans trust predictable robots more and view them more positively than less predictable robots [27, 71]. Thus, by improving the predictability of robots, we can improve the fluency and cohesion of human-robot teams.

Humans also trust agents more when we find them predictable [27]. With traditional controllers, however, there are no guarantees of pattern or regularity, so humans' mental models of robot teammates are often incorrect or incomplete [65, 9].

2.0.3 Mental Modeling

When humans collaborate, we build mental models of our teammates [122, 56, 125, 79]. Mental models are knowledge structures that help people to describe, explain, and predict events in our environment [108, 122]. Mental models help us navigate environments, make decisions, and reason about our collaborators. Human teams are so astute, they can even create shared mental models for the team, creating collective knowledge and expectations that lead to greater success [79, 56, 125]. Evidence shows that humans also build mental models of robots and derive their expectations of said robots from mental models [108, 90, 108, 11, 9]. In line with human factors research, work within human-robot interaction indicates that when humans and robots can build accurate mental models of each other, human-robot collaboration is more likely to be successful [90, 56, 125, 79].

2.0.4 Human Cognition

Humans reason in a fundamentally different, and often contradictory ways to our robot teammates. Artificially intelligent agents, embodied or otherwise, are built to optimize, but humans do not optimize when we plan or make decisions [40]. We satisfice—meaning we find a "good enough" solution to the problem [40, 3, 78]. People employ a variety of cognitive tools to do this, from using heuristics to pattern recognition [85, 1, 114]. Satisficing is not a weakness of human cognition; to the contrary, heuristic usage approaches rationality over the long term, and our brains developed it to navigate our environment, where optimization is computationally intractable [3, 45, 2, 114]. In human-robot teams, robot teammates are working on identifying and achieving the optimal solution for a specific set of parameters, whereas human teammates are agreeing upon a "good enough" solution, and these solutions are rarely the same.

The human brain floods with dopamine upon recognizing a pattern, thus, humans are strongly incentivized to find them [68]. Some scientists even consider pattern recognition and reasoning to be a cornerstone of higher intelligence [80]. As the human brain is wired for pattern recognition, actions that are pattern-based are more likely to be recognizable to human participants.

2.0.5 Patterning in Human-Robot Teams

There is strong evidence in the literature that using human cognitive tools within human-robot and human-agent collaboration can be highly effective [63]. Work has shown that human partners can learn conventions developed by artificially intelligent agents [102]. Significant work has also been done to integrate social conventions into collaborative agents [67, 106]. When robots explicitly adhere to human navigation conventions, humans find them more predictable and likeable, and the robot's ability to navigate is not compromised [92, 96]. Further, having people rely on conventions that they create themselves [20] or are already familiar with, such as "pinch" and "pull" motions that people use on their smartphones [41], leads to improvement in their ability to collaborate with a robot. Failing to account for human cognitive tendencies may obscure results,

and limit future work [9].

Overall, prior work in human–robot interaction and teaming has established the importance of coordination, transparency, and shared understanding, but often treats structure as fixed or secondary. Emerging research suggests that how robot behavior is organized, particularly through patterns or grouping, can meaningfully shape human prediction and interaction, especially as task complexity increases. These gaps motivate a closer examination of structure as a design dimension, which this thesis addresses by exploring how patterning can support more effective human–robot teams.

2.0.6 Multi-Agent Settings

Multi-agent HRI introduces challenges that extend beyond those encountered in single-robot settings. Rather than reasoning about the behavior of an individual system, humans must interpret, monitor, and predict the joint behavior of many agents operating simultaneously[12]. As the number of agents increases, the interaction becomes less about discrete entities and more about understanding collective dynamics, coordination patterns, and emergent behaviors[12]. This shift fundamentally changes the nature of the task, placing greater demands on perception, attention, and prediction, and raising questions about how system design can better support human understanding at scale[32, 74].

A central issue in multi-agent settings is the rapid growth of complexity with increasing agent count[74]. Prior work has shown that human performance in tracking and predicting multiple moving entities degrades as density and dynamism increase, reflecting well-established limits in attention and working memory[12]. Importantly, this degradation is not solely a function of the number of agents, but of how their behavior is structured. When agents act independently, the cognitive burden scales poorly, requiring observers to track many parallel trajectories[12]. This has led to a growing recognition that effective complexity; the perceived or cognitively relevant complexity of the system, may be more important than raw agent count in determining human performance[12].

To address these challenges, existing approaches in multi-agent HRI have largely focused on improving observability or constraining behavior[32, 74]. Visualization and interface design techniques aim to externalize information through trajectory overlays, intent indicators, or other augmentations, helping users access otherwise implicit system states. Alternatively, behavior-level interventions impose coordination structures such as formations, lanes, or shared control policies that regularize agent motion and improve predictability[117, 104]. While these approaches can improve performance, they often operate either by adding information, risking visual or cognitive overload, or by restricting agent autonomy in ways that may limit flexibility or realism.

2.0.7 Collective Emergent Behavior

Emergent behavior refers to complex, coordinated patterns that arise from the local interactions of individual agents following relatively simple rules[25]. In crowd dynamics, these behaviors—such as spontaneous lane formation, collective turning, and synchronized flow through bottlenecks do not require explicit global control or communication. Instead, they emerge organically as each individual continuously adapts to the movements and intentions of others. This bottom-up organization is a hallmark of many natural and social systems, highlighting how structure and predictability can result from decentralized decision-making[25].

Emergent behavior is closely tied to human perceptual and cognitive abilities to detect and respond to patterns in their surroundings [69, 25, 48]. As pedestrians navigate shared environments, they continuously interpret subtle cues, such as speed, gaze direction, and interpersonal spacing, to infer others' intentions and adjust their own motion accordingly [48]. This innate patterning ability allows individuals to synchronize movements without explicit communication, reinforcing collective structures like walking lanes and flow partitions. In turn, these emergent patterns reduce cognitive load and enhance overall efficiency by creating predictable pathways through dynamic crowds [48]. The interplay between human pattern recognition and emergent behavior underscores the importance of designing autonomous systems that are not only optimized but also engage with the implicit social patterns that guide human group motion.

2.0.8 Social Force Models

Social force models constitute a widely adopted theoretical framework for representing pedestrian and crowd dynamics through mathematically defined "forces" governing agent motion[48]. In this formulation, each pedestrian is treated as a self-driven entity with a preferred velocity toward a goal, while additional repulsive and sometimes attractive forces modulate interactions with other agents and the environment. These forces are not literal physical interactions but abstractions that encode behavioral tendencies such as maintaining personal space, avoiding collisions, and adhering to spatial conventions. By embedding these behavioral principles into continuous optimization of trajectory and velocity, social force models enable systematic simulation and analysis of navigation behaviors at both local and global scales.

Crucially, social force models facilitate the emergence of complex collective phenomena from simple interaction rules. Empirical studies have demonstrated their ability to replicate observed pedestrian behaviors, including lane formation in bidirectional flows, group cohesion among socially connected individuals, and non-linear congestion patterns in constrained environments[48, 69]. Extensions to the foundational model further incorporate elements such as predictive collision avoidance, heterogeneous population characteristics, and context-specific behaviors relevant in emergency or high-stress scenarios [112, 48, 13, 8, 87, 29, 70, 26].

In particular, social force models that explicitly incorporate group dynamics extend the foundational framework by accounting for the social bonds and shared goals that influence pedestrian behavior[88]. Rather than treating each individual as an isolated agent, these models introduce group cohesion and alignment forces that preserve proximity and coordinated movement among members of the same group [88]. Such forces capture underlying social motivations and ensure that group members adjust their trajectories in ways that maintain collective identity while still respecting collision avoidance constraints. This enriched structure allows simulations to reflect more realistic, heterogeneous populations where interactions vary significantly between intra-group and inter-group encounters[112, 124, 121].

These group-aware extensions are particularly important for understanding how collective behavior shapes crowd-level phenomena. For example, cohesive groups tend to move more slowly and occupy more space, creating flow disturbances and bottlenecks in tightly constrained environments [112, 124, 121]. Group models can also capture phenomena such as group splitting and merging, leadership dynamics, and the influence of strong social ties during evacuation scenarios [112]. In applied contexts such as autonomous navigation and human-robot interaction, leveraging group-sensitive social force models enables robots to interpret and respond to group formations more appropriately, whether by integrating into existing clusters or avoiding disruption of social units [124]. In this way, group-oriented social force frameworks provide a more nuanced and human-centered foundation for analyzing and shaping motion in shared environments.

Chapter 3

Generating Pattern-Based Conventions for Predictable Planning in Human-Robot Collaboration

The primary goal of this first work is to establish a clear and operational definition of what constitutes a pattern in the context of human-robot teaming. To begin, we will focus on establishing a pattern definition for the subtask-planning level. This includes specifying the structural features that define a pattern, the conditions under which it is considered to be present, and how it can be represented in a way that is both measurable and reproducible. In parallel, we also need to determine whether the use of patterns has any meaningful effect on humans who work with them. Before optimizing or refining a pattern-based approach, it is important to verify that pattern usage produces measurable differences relative to a baseline without patterns.

In addition to establishing that patterns have an effect on humans who work with robots, we must also develop a robust way to evaluate their quality. This requires defining metrics that capture whether a pattern is deterministic or unique given the environment, rather than simply whether it exists. These metrics should allow us to compare different patterns and assess their performance. Finally, this proposed approach must be validated via an appropriately structured, IRB-approved study with a human and robot.

3.1 Introduction

Human difficulty with accurately modeling and predicting robot behaviors prevents the integration of robots into human-populated environments. Prior work indicates that the more ef-

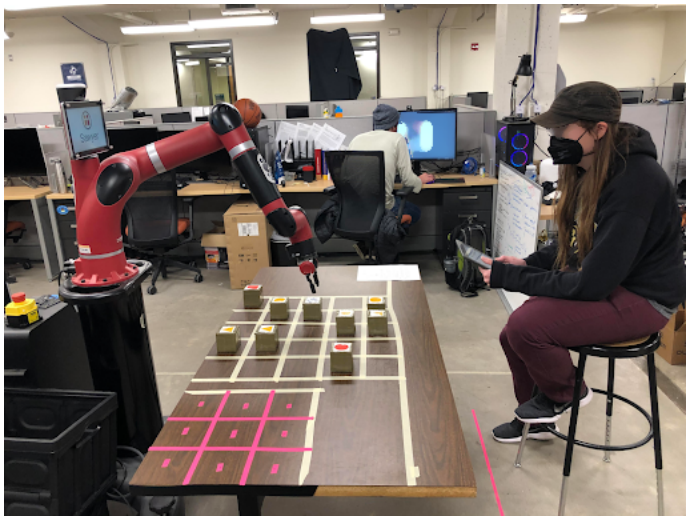

Figure 3.1: A participant plays a collaborative block-selection game with a robot. By using PACT to augment its planner, the robot's actions are more predictable to the participant over multiple episodes and multi-task time horizons, and the robot is viewed as a better teammate.

fectively a human can model their robot teammate, the better the team will be able to perform [90]. However, humans struggle to build accurate and effective models of robots [65, 9] and often find them unpredictable even in very simple environments [6]. This limits team performance, as humans prefer to work with agents they find predictable and trust unpredictable agents less [27].

As humans struggle to predict agent behavior, agents are simultaneously attempting to predict and adapt to humans. Prior work has been done to improve an agent's ability to predict human actions [98, 39, 42, 76] as well as adapt to human behaviors [17, 47, 106]. However, for collaboration to succeed, both human and agent need to be mutually predictable, and there are significant technical gaps in improving humans' ability to predict agents' actions.

In contrast to human-robot teams, human-human teams are extremely skilled at collaborative tasks where synchronization, coordination, and prediction of each other's behavior is necessary—such as assembling a bookshelf or making a football pass. Part of this gap in performance can be explained by the distinct sets of tools that humans and robots each use to accomplish tasks. Humans do not often rely on optimization as a cognitive tool, and instead use heuristics and pattern recognition [40]. Notably, humans have difficulty in predicting what the robot will do next as the

robot is not using cognitive tools the human is familiar with.

The cognitive processes humans rely on, while simpler than optimization methods, are more adaptable than and often outperform such techniques, especially in complex environments where optimization is computationally intractable [3, 40, 85, 45]. One of the cognitive tools that humans employ is the ability to identify and process patterns, such as recognizing rhythm in a song or the color order of changing stoplights. Pattern recognition has developed via our evolution as a species and is facilitated by specific structures in our brains [80]. Even preschoolers are capable of duplicating, extending, and abstracting patterns to new environments [95]. These are deeply ingrained cognitive processes that humans are adept at using.

Within the context of teaming, humans extensively rely on conventions in order to effectively coordinate behavior. Conventions are a form of shared knowledge [102] that teammates can use in collaborative tasks to synchronize their actions. Patterns—a predictable sequence [95] of actions—can make conventions easier to learn and follow. For a pattern-based convention to be meaningful to a human, the pattern must be human-perceptible, i.e., based on features that a human can observe. Using pattern-based conventions leverages innate human pattern-processing abilities, making the conventions intuitive for people to identify [80] and predict. In order to facilitate human-robot collaboration, in this work we propose a filtering algorithm which enables an embodied agent to set conventions for the team by using a human-perceptible pattern to restrict its actions in a given situation to a more predictable set. We refer to this algorithm as Pattern-Aware Convention-setting for Teaming (PACT).

PACT can select patterns of varying complexity depending on the context, and is robust to changes in the environment. A pattern selected by PACT can continue to be used without alteration even if the task space changes over time. **Our results show that by using PACT, not only does the robot become more predictable to its human teammates, but team performance as well as perceptions of the robot improve.** These findings are supportive of the hypothesis that, by leaning into familiar cognitive processes, humans can more readily identify the robot's intentions, understand how the robot is making decisions, and abstract the robot's

behavior into new environments accurately. Not only is PACT effective, it demonstrates the benefit in adapting human cognitive strengths for robots to use during collaboration.

3.2 Background and Related Work

When humans collaborate, we build mental models of our teammates [122]. Mental models are knowledge structures that help people to describe, explain, and predict events in our environment [108]. Human teams are so astute, they can even create shared mental models for the team, creating shared knowledge and expectations that lead to greater success [79]. Evidence shows that humans also build mental models of robots [108]. In line with human factors research, work within human-robot interaction indicates that when humans and robots can build accurate mental models of each other, human-robot collaboration is more likely to be successful [90, 56, 125, 79]. Humans also trust agents more when we find them predictable [27]. With traditional controllers, however, there are no guarantees of pattern or regularity, so humans' mental models of robot teammates are often incorrect or incomplete [65, 9].

Humans reason in a fundamentally different, and often contradictory way to our robot teammates. Artificially intelligent agents, embodied or otherwise, are built to optimize, but humans do not optimize when we plan or make decisions [40]. We satisfice—meaning we find a "good enough" solution to the problem [40, 3, 78]. People employ a variety of cognitive tools to do this, from using heuristics to pattern recognition [85, 1, 114]. Satisficing is not a weakness of human cognition; to the contrary, heuristic usage approaches rationality over the long term, and our brains developed it to navigate our environment, where optimization is computationally intractable [3, 45, 2, 114]. In human-robot teams, robot teammates are working on identifying and achieving the optimal solution for a specific set of parameters, whereas human teammates are agreeing upon a "good enough" solution, and these solutions are rarely the same.

Much of the recent work in human-robot collaboration focuses exclusively on improving the performance of the robotic agent. Works that attempt to predict human actions or their path directly have seen success within the environments they tested in [98, 39, 42]. There has also

been a significant effort to adapt successful methods in competitive environments to collaborative environments [58, 17, 111, 52], though this is very difficult. Approaches that are highly effective in competitive environments are challenging to adapt to collaborative environments [51]. What makes many self-play approaches successful—a policy that is convoluted and difficult for opponents to counter—is a drawback in collaborative settings. What results is a large drop in performance when trained agents are tested with humans rather than other agents [106, 58]. These approaches also do not capture the full scope of human collaboration within their environments [51].

There is strong evidence in the literature that using human cognitive tools within human-robot and human-agent collaboration can be highly effective [63]. Work has shown that human partners can learn conventions developed by artificially intelligent agents [102]. Significant work has also been done to integrate social conventions into collaborative agents [67, 106]. When robots explicitly adhere to human navigation conventions, humans find them more predictable and likeable, and the robot's ability to navigate is not compromised [92, 96]. Further, having people rely on conventions that they create themselves [20] or are already familiar with, such as "pinch" and "pull" motions that people use on their smartphones [41], leads to improvement in their ability to collaborate with a robot.

Failing to account for human cognitive tendencies may obscure results, and limit future work [9]. One such cognitive tendency is pattern recognition. The human brain floods with dopamine upon recognizing a pattern, thus, humans are strongly incentivized to find them [68]. Some scientists even consider pattern recognition and reasoning to be a cornerstone of higher intelligence [80]. As the human brain is wired for pattern recognition, actions that are pattern-based are more likely to be recognizable to human participants. By using PACT, we can select patterns that are as recognizable as possible.

3.3 A Framework for Patterns-Based Conventions

In this section we detail PACT, an entropy-based algorithm to select the most appropriate pattern to use in a given environment. The central cognitive science concept that underpins this

approach is the human tendency toward pattern recognition and usage. By playing into known strengths of human cognition, the robot's behavior becomes more recognizable, predictable, and understandable to human teammates.

3.3.1 Definitions

PACT takes the tuple $\{T, F, r\}$ as input to determine the ideal pattern for a particular task space, where:

- $T = \{t_1, t_2, ..., t_n\}$ is the finite set of subtasks an agent must complete. T is unordered, but subtasks within T may have ordering constraints imposed by prerequisites.
- $F = \{f_1, f_2, ..., f_m\}$ is a set of functions that map from a subtask to a feature of that subtask. (e.g., $f_1(t) \rightarrow$ "$circle$", $f_2(t) \rightarrow$ "red")
    - $f_i = [v_1, v_2, ..., v_k]$ is a feature vector representing a characteristic (e.g., for a feature "color" there may be categorical values { "red", "green", "blue"} encoded as a one-hot vector. A color feature could also be represented as a continuous three-dimensional vector of RGB values).
- A **Rule** is a function that sorts subtasks in T using a comparator function over output from one or more features in F.
- r is the maximum number of Rules that PACT is allowed to combine to form a **Pattern**, a hyperparameter selected by the user prior to Pattern formation.
- A **Pattern** is an ordered sequence of between 1 and r Rules that augments available subtasks in T for a planner to select from in a given state. Rules are applied sequentially to filter out or augment the cost of elements in T to inform plan generation.

Figure 3.2: In this illustration of the PACT algorithm, we use a scenario in which a natural disaster has occurred in a coastal town. Critical infrastructure must be checked for damage, and an autonomous drone as well as a human team on the ground are tasked with damage assessment. In this time-sensitive task, communication between the drone and humans is limited. Each location indicated on the map has features used by PACT: whether the location contains humans that sheltered in place (red), the type of infrastructure (blue), and the likelihood the location is flooded (green). This scenario does not consider the distance traveled by the drone to be a constraint, but such constraints can easily be added to guide PACT's pattern-selection.

3.3.2 Rule Formation and Application

A Rule is a data structure that contains a sorting function and a set of features to apply it to. Given a set of subtasks, a Rule filters it down to a subset of subtasks that the agent can perform (while still being consistent with the Rule). For example Figure 3.2 shows an environment in which an autonomous drone must check critical infrastructure after a natural disaster. Communications are down, so the drone is unable to communicate reliably with a human ground crew, making the predictability of the robot critical. Here, the set of subtasks T is the set of locations the drone must check and document. Each location has three features: the estimated flooding risk (which we discretize into low, medium, and high risk), the type of critical infrastructure (police station, power substation, water treatment plant, and hospital), and whether or not human staff sheltered in place there. A Rule based on the presence of humans could be ["no humans", "humans"], such that the robot would visit all places without humans sheltering, followed by those locations with humans. Another Rule based on the flood risk could be ["high", "medium", "low"]. Applying the flood risk Rule ["high", "medium", "low"] to the locations in T would result in filtering the locations down to the subset of locations with "high" flood risk. Each location in this subset would be visited by the

drone. Then, with no more "high" flood risk locations, the drone would visit all "medium" flood risk locations, and so on. For this Feature, because each location has one of three possible values, there are 3! possible orderings, meaning this Feature (flood risk) has 3! possible Rules that could leverage it. Thus, given a set of n categorical Features, where each Feature i has k_i possible values, there are at most $\sum_{i=1}^{n} k_i!$ single-Feature Rules that can be generated. For Rules over features with non-categorical values, the space of orderings is technically infinite and depends entirely on how complex the comparator function encoded in the Rule is, but by imposing a restriction to sort values in either an ascending or descending order we may assume two Rules per continuous feature. In the drone example given by Figures 3.2 and 3.3 there are 32 possible Rules that can be used. Without loss of generality, in this work, each Rule is generated from a single feature.

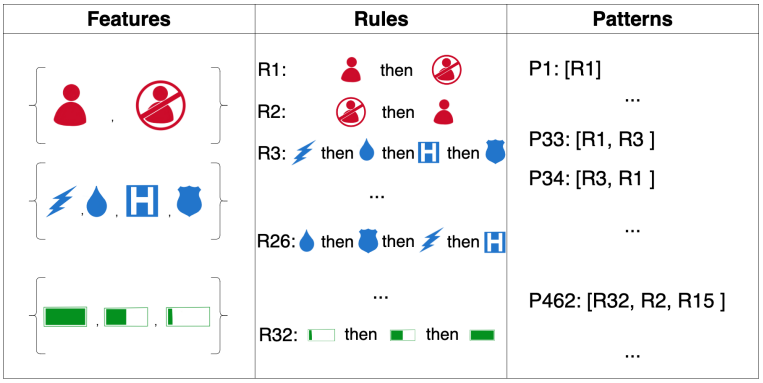

Figure 3.3: In this example, we describe the formation of Rules and Patterns from the scenario in Figure 3.2. The left column shows the three features of each location we are using (human presence, infrastructure type, and flood risk), and the possible values for each feature. The center column shows the ways Rules can be constructed from features by imposing an ordering on the possible values of a feature. The right column shows how a Pattern is constructed by applying one or more Rules. Note that Patterns may not have conflicting Rules; we choose at most one rule per feature.

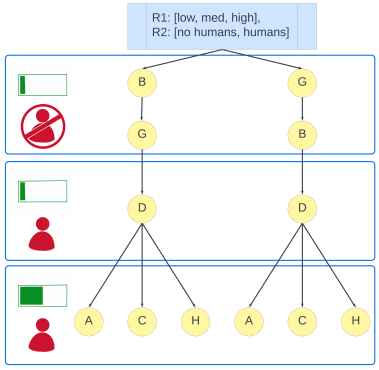

Figure 3.4: For each Pattern, a Pattern Tree is constructed to easily identify all allowable orderings of subtasks. In this Pattern, the first subtasks are those locations that are low flood risk and have no humans sheltering in place (B,G). The second allowable subtasks are the remaining low risk, no human locations, which are appended to the tree. All allowable orderings of length 2 can thus be obtained by traversing the tree to depth 2. For the third subtask, there are no remaining low risk locations with no humans, so the low risk locations with humans are selected (D). There are no medium flood risk locations that do not have humans, so the fourth possible subtasks are those that are of medium risk with humans. The tree is constructed in this manner until it reaches a depth equal to the number of subtasks.

3.3.3 Pattern Formation and Application

The hyperparameter r is set by the user prior to the generation of Patterns in order to determine the maximum allowable complexity for the Patterns. r can be at most equal to the number of Features and must be at least one. With a larger r, Patterns can be more complex and are thus more likely to be able to impose a fully deterministic ordering of tasks in a plan, but this increase in complexity may also make the Pattern too difficult for a human partner to identify and follow.

A Pattern is a data structure that contains a sequence of between 1 and r Rules. Given a set of subtasks, the Pattern determines the subset of next possible subtasks. The initial set of subtasks is passed to the first Rule in the sequence, which returns the subset of allowable subtasks.

This subset is passed to the second Rule in the sequence, continuing through the full sequence of Rules to obtain the final subset of possible subtasks for the given Pattern. Figure 3.4 illustrates the Pattern [“low”, “medium”, “high”], [“no humans”, “humans”]. First, the set of locations the drone must visit is filtered down according to the first Rule (flood risk), leaving just the locations with “low” flood risk. This subset of locations is then passed on to the second Rule (human presence) to be filtered down to locations with “no humans”. The drone will have to visit all locations in this subset (B and G) in any order before moving on to locations with different values for these features. After these locations, pictured in the first box of Figure 3.4, are checked, the remaining locations are passed to the first Rule and then the second to obtain the subset of locations that are “low” risk with “humans” (location D). After this location, the Pattern filters down to an empty set, as there are no locations with “medium” flood risk and “no humans”. The Pattern will then identify locations with “medium” flood risk and “humans”, which will be visited before locations with “high” risk and “no humans” (F) and then locations with “high” risk and “humans” (E).

3.3.4 Pattern Trees

Calculating a score to evaluate the effectiveness of a given Pattern requires evaluating the possible orderings of subtasks that it imposes throughout the plan it generates (or a sampled subset if otherwise infeasible). To efficiently compute and organize this for each Pattern, we construct a Pattern Tree. An illustration of a portion of a Pattern Tree generated from the drone example can be seen in Figure 3.4. The first level of the tree is determined by the possible first subtasks given the Pattern and T. For each subsequent level, the children of a node are determined by assuming the path from root to parent node specifies the sequence being followed, and applying the Pattern to the remaining subtasks. The final tree will have $|T|$ levels, as the entire sequence will be generated. Thus, traversing to the ith level of the tree will reveal all possible subsequences of length i for a given Pattern. This simplifies Pattern evaluation calculations as matching subsequences of length $i-1$ can be obtained for all Patterns quickly, and all possible ith subtasks can also be obtained by indexing the children of all nodes in the $(i-1)$th level of the tree. For tasks with prohibitively

large amounts of subtasks, Monte Carlo methods can be applied to approximate the Pattern Tree.

3.3.5 Pattern Scoring Metric

To determine the most appropriate Pattern for a given T, we propose a scoring metric that can be applied to a set of possible Patterns (which we refer to as the **Pattern Bank**). Patterns with lower scores are more preferable. We define this score (λ) for a given Pattern (p) as:

$$\lambda_p = \sum_{i=1}^{|T|} \mathcal{H}(T_{i,p}) + \left(\frac{|P_{i,\text{shared}}| - 1}{|P|} \right) * \mathcal{H}(T_{i,\text{shared}}) \tag{3.1}$$

Where:

- p is the Pattern for which the score is being calculated.
- T is the set of subtasks the agent must perform.
- $\mathcal{H}(x)$ is an entropy calculation for the collection x.
- $T_{i,p}$ is the collection of all possible subtasks at the ith step of planning given the Pattern p.

This can be extended over sets of Patterns as follows:

- P is the set of possible Patterns given $\{F, r\}$.
- $P_{i,\text{shared}}$ is the set of Patterns that share at least one possible sequence of length $i-1$ with the given Pattern p.
- $T_{i,\text{shared}}$ is the collection of all possible subtasks at the ith step for all Patterns in $P_{i,\text{shared}}$.
- All sequences of length $i-1$ are allowable.

This scoring metric allows for selection of a pattern that is both as deterministic as possible (first term) as well as unique (second term). Favorable patterns are those that become unique in their possible sequences as soon as possible (easier to identify/ legible [31]), while also being as deterministic as possible (easier to follow).

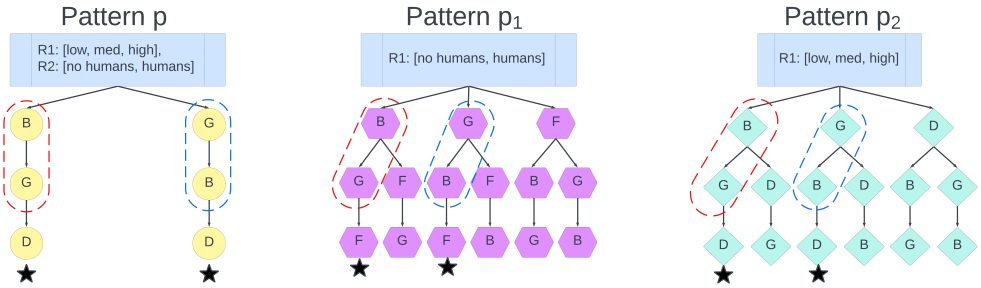

Figure 3.5: This figure illustrates how the second term of the score in Eq. 3.1 is calculated for a given Pattern (the same Pattern shown in Figure 3.4) when $i = 3$. First, the possible orderings of length $i-1$ are identified for the given Pattern, seen in the left tree. There are two possible subtask orderings of length $i-1$, highlighted in red and blue. However, these orderings are not unique to this Pattern. There may be other Patterns in the Pattern Bank that share these orderings of length $i-1$. Two such Patterns are shown here, with the matching orderings circled. If a human partner observes the robot going to B then G, they cannot distinguish between the Pattern the robot is following and these other Patterns. The (starred) children of these shared orderings are extracted from all Patterns in the Pattern Bank, and the entropy over this group is calculated. For this group of three trees, the group would be (D, D, F, F, D, D).

3.3.6 PACT

While the algorithm can be viewed in its entirety in pseudocode within Algorithm 1 in the appendix, we provide an intuitive walkthrough here for ease of understanding. Prior to applying PACT, we create a Pattern Tree for each Pattern in the Pattern Bank. We then initialize data structures that keep track of the best patterns and their scores. For each Pattern (p) in the Pattern Bank we calculate a pattern score, weighing how deterministic the pattern is and how much overlap in resulting plans there is with those generated by other Patterns in the Pattern Bank (i.e. how unique the pattern is compared to others). A pattern score is a summation of subscores calculated for the selection of each subtask in the sequence imposed by the Pattern. Scores start at zero and increase at each step. The subscores have terms related to entropy (i.e., how deterministic the plan imposed by the Pattern is) and uniqueness (to bias against Patterns that generate plans that can be explained by other Patterns).

The first term of Equation 3.1 is the entropy over the distribution of ith possible subtasks when a sequence of subtasks is being constructed using the given Pattern p. When using p to order

the subtasks, the allowable sequences can be determined by traversing the tree. Thus, the subtasks that could be ith in a sequence that conforms to p are all those nodes at a depth of i in the Pattern Tree. In Figure 3.4, when $i = 3$, the nodes used for this calculation are in the middle box (depth of 3). When $i = 4$, the nodes used for this calculation are those in the bottom box. The first term for i is the calculated entropy for the set.

Figure 3.5 illustrates the calculation for the second term when $i = 3$. This calculates how unique p is, i.e. how much overlap there is between p and other Patterns in the Pattern Bank. The term is composed of an entropy value and a discount.

When determining how unique an ordering induced by p is, the possible orderings of subtasks must be compared with those of other possible Patterns. In Figure 3.5, p is the Pattern on the far left, with circular nodes. When $i = 3$, there are only two possible subtask orderings of length 2 that follow the Pattern, circled in red and blue dashed lines. If the robot is using p to order its subtasks, a human partner will observe one of the circled orderings. However, these orderings may also comply with other Patterns within the Pattern Bank. The trees with hexagonal and diamond nodes in Figure 3.5 are other Patterns in the Pattern Bank which have some orderings of length $i - 1$ (2) in common with the target Pattern p—also circled with dashed lines. If the robot goes to location B then on to G, this behavior can be explained by p, but also by these other Patterns, which may lead to the human partner to believe the robot is following a Pattern other than p leading to confusion or difficulty predicting the robot in the future, as their mental model of the robot is incorrect. The second term identifies the children of these shared sequences, marked in the figure with stars, and calculates the entropy over them. Thus, Patterns that produce sequences of subtasks that are unique have lower scores, and Patterns that produce orderings of subtasks that are shared across many Patterns have higher (worse) scores.

This entropy calculation is then discounted by the proportion of the Pattern Bank that has an ordering of length $i - 1$ in common with p. This is done to penalize Patterns that could be mistaken for a greater number of other Patterns. If p shares many orderings of length $i - 1$ with one other Pattern, this will lead to less confusion on the part of a human partner than if p shares

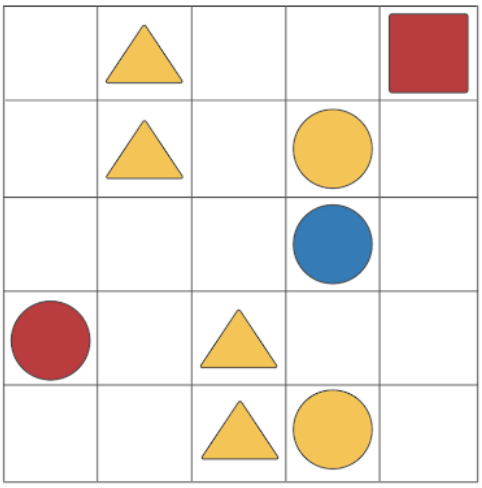

Figure 3.6: The layout for the collaborative game. Nine blocks, each with a color, shape, and reward value are placed on the grid. Only the robot has knowledge of the rewards. Both players secretly select a block by color and shape, and if they coordinate, the robot removes a block from the grid.

a few orderings of length $i - 1$ with many Patterns in the Pattern Bank.

When all of the subscores have been calculated and summed, we compare the total score for the Pattern to the minimum score, and store all minimum-scoring Patterns. When we have scored all Patterns, we return every minimum-scoring Pattern for subsequent selection and use by the planner.

Pattern scoring and selection is performed offline, done before the robot engages with an environment. While the Pattern can be updated or changed, a Pattern deemed to be the most suitable for a target set of environments can and should continue to be used in other environments the robot acts in to maximize predictability, as long as the features used in the Pattern remain present in these other deployment environments. Changes made to the Pattern during interaction with humans may make the robot less predictable, and this work promotes the use of one Pattern kept consistent even when the robot finds itself in previously unseen deployment environments.

3.4 Experimental Evaluation

PACT can be applied to any planning problem for which the overarching task can be decomposed into a predefined set of subtasks or goals (e.g., search a set of ten locations for survivors).

PACT can be applied to situations where the robot is working with one or more humans, such as remote sample recovery, wherein PACT would make it easier to predict where the robot would be retrieving samples from, allowing humans to parallelize efforts by focusing on areas that the robot is not or to assist robots by traveling to their next destination without explicit communication requirements. PACT may also be used in scenarios where human and robot are simply sharing a workspace, where increased predictability of which object the robot will grab next allows humans to more easily navigate around or more safely work with the robot.

However, in these scenarios as well as in other more complex environments, there is a significant amount of extraneous side-channel information that people may use to predict the robot's behavior. People may wait for several moments to determine where the robot is headed next, take time to simply observe the robot, or even be provided with information from the robot itself. In order to effectively test PACT, and to show that the planner's order of subtasks alone is driving increased predictability, all of this information must be removed. Any effective testbed for such a system must be framed as a coordination problem, so that the human does not have the opportunity to observe the robot without taking any action themselves. The coordination task is structured such that only by accurately predicting the robot's actions can the team succeed, and there is no other information the human can rely on other than previous robot actions and their own mental model of the robot.

PACT can be applied to a broad range of planning problems that can be constructed from this maximally constrained coordination problem, by relaxing the testbed's requirements of forced simultaneous action selection or inability to wait and observe the robot. While this makes the task of coordinating with and predicting the behavior of the robot significantly more difficult, it allows for a stronger assessment of the effectiveness of PACT than would occur in a more realistic

collaborative scenario with additional side-channel information available.

Thus, we evaluate the efficacy of PACT through a collaborative game involving a human and robot that rewards teams whose members' task selections are predictable.

3.4.1 Game Environment

The collaborative game is played on a five by five grid on a table in a shared workspace (Figure 3.6). At the beginning of a round, nine blocks are placed in unique locations on the grid. Every block is assigned a unique numerical value between one and nine, which is neither known nor observable by the human participant and is used to calculate the score for a successful move, representing the reward function that a traditional robot planner would attempt to maximize. At the start of each turn, the participant and Sawyer, a 7-degree-of-freedom robotic arm, both select a block without seeing the choice of their teammate. Participants make their selection on a tablet, and are allowed to select any block type (e.g., "blue triangle"). When both teammates have made their selection, Sawyer reveals its selection on its screen, and the participant receives an update on the tablet showing both players' selections as well as score information. If the team members each select blocks with fully matching visual features (e.g., both with yellow circles on them), Sawyer removes one block that matches those features from the grid.

The team receives a positive reward based upon the numerical values of the blocks remaining and the number of blocks the team had to choose from; as the number of blocks on the board decreases (and it is more likely that teammates could coordinate by chance), the reward decreases. The game is scored as follows:

$$S(t) = \begin{cases} -10t/(n+1) & n \text{ matching features} \\ B_{sum} + 5(|B_{rem}| + 1) & \text{all features match} \end{cases}$$

where:

- t = current turn number

- B_{rem} = set of blocks remaining on the board
- $B_{sum} = \sum_{i=0}^{|B_{rem}|-1} B_{rem}[i].numeric\_value$

If the team does not agree on the same type of block, a penalty is assessed to the team. The size of the penalty increases as the game progresses, so an inability to coordinate early in the game is not penalized as harshly as failing to coordinate on the last few blocks. If the team is able to coordinate on a subset of features, i.e., both players select the same color or shape (but not both), the penalty assessed is reduced; teams that are able to coordinate along some axes are not penalized as much as teams that cannot coordinate on any features. Teams must coordinate to remove all nine of the blocks from the grid to complete a round.

It is important to note that the sequence of blocks that the robot will select is determined prior to the first turn: **this experimental setup is designed to test human understanding and prediction of robot behavior, not robot adaptation to human behavior**. Regardless of what the participant selects, the robot will always select the next block in the predetermined sequence.

This means that the robot will continue to select the same block until the participant matches the robot's selection. Upon the completion of each round, a new set of nine blocks is placed in the workspace in a new configuration. The numerical value of each block is also new, with no relationship between the value and any of the human-visible features. In other words, the reward function changes with each episode and is never shown or explained to the human teammate. This design decision illustrates the trade-offs between capability (reward maximization) and predictability (pattern adherence) when coordinating as a human-robot team.

3.4.2 Applying PACT to the Coordination Domain

The variables required to utilize PACT are defined as follows:

- T = A set of subtasks, one for each of the nine blocks in the workspace.

- F = A set of functions mapping each block subtask (by unique id) to values of features of the blocks— "color": {"blue", "red", "yellow"}, "shape": {"circle", "square", "triangle"}, "position": {"row":{1, 2, ..., 5}, "column": {1, 2, ..., 5}}
- $r = 2$, such that only up to two Rules may be ordered to create a Pattern

The **position** Rules ordered the values in either ascending or descending order, operating over either only a single field (either "row" or "column") or both (e.g., rows descending and columns ascending). In our environment—because the human and robot select blocks by color and shape on the interface (e.g., a "yellow triangle block", without specifying location)—when calculating entropy over subsequent tasks to select, all tasks involving blocks of the same shape and color are treated equally regardless of position.

3.4.3 Experimental Design

Study participants ($n = 28$) were assigned randomly into one of three conditions in a between-subjects design:

- Reward-Maximizing: The participant works with a robot that selects blocks in the order that will maximize the team score in the event of perfect coordination, analogous to traditional reward optimization approaches.
- PACT Pattern: Participants are on a team with a robot that selects blocks following a pattern-based convention, generated and selected by PACT such that the pattern score is best for the set of tasks to be completed in the first round environment.
- Median Pattern: Participants work with a robot that selects blocks by following a pattern that achieved a median score when compared against all possible patterns in the first round environment.

Patterns selected for the PACT and Median groups are based on the first round environment and remain the same for all subsequent rounds of gameplay, despite environment changes. This allows

us to evaluate team performance in both a 'target' environment that may be known and optimized against in advance (first round) and in new environments not explicitly optimized for (subsequent rounds).

3.4.4 Study Protocol

Consent was obtained from all participants, preceded by a brief check of participants' ability to distinguish between the block colors. One participant self-identified as colorblind, though not a form of colorblindness that would prevent them from distinguishing between the colors used. Participants were then given a randomly generated six-digit identifier to link their survey responses, and were randomly assigned to an experimental group. Following this, participants filled out a pre-experiment survey about their experience with robots, attitudes about robots, and initial sentiments toward Sawyer, the robot used in the experiment. Experimenters then explained the collaborative game, answered questions, and participants began gameplay. After each round of the game, participants answered questions about their cognitive fatigue, ability to predict the robot's behavior, and confidence in their team. After three rounds of the game, a third type of survey was administered. Participants were shown five novel game set ups, and were asked to identify which color and shape block the robot would select first and last in each given game. Participants were also given the option to mark that they were uncertain about either feature. Finally, a post-experiment survey was conducted, again surveying participants about their sentiments about the robot, their game comprehension, as well as questions about the team dynamics and performance of each team member. Following the completion of the survey, participants participated in a brief unstructured interview and debrief. The duration of the experiment was approximately sixty minutes.

3.4.5 Measurement

28 participants were recruited from the student community of our university for the IRB-approved human subjects study. Pre-experiment survey questions were taken from NARS, RoSAS, and previous HRI work [91, 16, 99]. Between rounds, participants answered selected questions

from the NASA Task Load Index [44] to measure their cognitive fatigue and frustration, as well as several questions about their confidence in their choices. A "Round 4" survey consisting of five novel game setups was created specifically for this experiment in order to measure participants' ability to abstract the robot's behavior into a new environment. The post-experiment survey consisted of questions from RoSAS, identical to those asked in the pre-experiment survey, survey questions about the fluency of the team [**?**], as well as custom questions adapted from the team fluency questions.

3.4.6 Hypotheses

We conducted an ethics board approved human-subjects study to investigate the following hypotheses regarding the effectiveness of PACT within a human-robot collaborative coordination task:

- H_1: Participants who work with the robot using PACT will have a more positive attitude about the dynamics of the team (i.e., coordination, mutual understanding, teamwork, etc) compared to all other groups.
- H_2: Participants who engage with the robot using PACT will have a more positive perception of the robot than participants in the Reward-Maximizing and Median Pattern groups.
- H_3: Constraining the robot's behavior to follow any patterns-based convention will result in better team performance on the task, as well as an improvement in participants' ability to predict the robot's actions.

3.5 Results and Discussion

Of the 28 individuals who participated, the data of one participant was excluded due to noncompliance with instructions. We did not observe any multimodalities within the data.

We found a significant effect from the PACT Pattern condition on participant perceptions of the team's dynamics, **validating $\mathbf{H_1}$**. Post-hoc comparisons using Tukey's HSD test (Figure 3.7),

indicate that participants felt that that robot picked the best block for the team during gameplay compared to the control condition of Reward-Maximizing ($p = 0.0353$) as well as the Median Pattern group ($p = 0.0493$). Additionally, PACT Pattern participants did not feel that swapping the robot out for a human teammate would result in better performance when compared to the Reward-Maximizing group ($p < 0.004$) as well as the Median Pattern group ($p < 0.009$), indicating that **PACT Pattern participants viewed the robot as performing at least as well as a human teammate would have**.

We also found a significant effect caused by the PACT Pattern condition on perceptions of team fluency, as indicated by Figure 3.8; the PACT Pattern condition resulted in a significantly higher perception of team fluency as compared to the Reward-Maximizing baseline ($p = 0.0178$), while there was no significant difference for participants in the Median Pattern group compared to the Reward-Maximizing condition. Additionally, there was a significant effect caused by the PACT Pattern condition on participants' perception of whether or not the participant and robot were good teammates to each other. PACT Pattern participants reported significantly more positive perceptions of themselves as a teammate to the robot when compared to the Reward-Maximizing ($p = 0.0129$) and the Median Pattern group ($p = 0.0336$), as well as whether the robot was a good teammate to the participant when compared to the Reward-Maximizing group($p = 0.0121$). **The PACT Pattern group was also the only pattern-based group that saw a significant difference over the Reward-Maximizing condition when asked if they would work with the robot again ($p = 0.0287$)**.

Post-hoc comparisons using Tukey's HSD test indicate a **partial confirmation of $\mathbf{H_2}$**. While there was a significant effect from the PACT Pattern treatment on the likeability of the robot when compared to the Median Pattern group ($p < 0.05$), there was no significant difference compared to the Reward-Maximizing group ($p = 0.1$), Figure 3.8.

We also found a significant effect from both pattern conditions on team performance, **validating $\mathbf{H_3}$**. Post-hoc comparisons using Tukey's HSD indicate significantly higher normalized scores for both the PACT Pattern and Median Pattern groups across all three rounds, as indicated

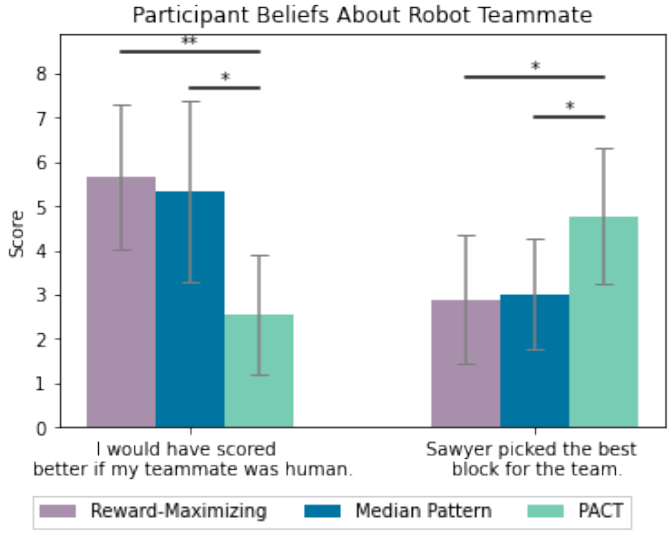

Figure 3.7: There were significant improvements in PACT Pattern participant belief that the robot selected the right block for the team over the Reward-Maximizing ($p = 0.0353$) and Median Pattern ($p = 0.0493$) groups, as well as if the participants believed a human partner would have led to greater success. (Reward-Maximizing $p < 0.004$, Median $p < 0.009$).

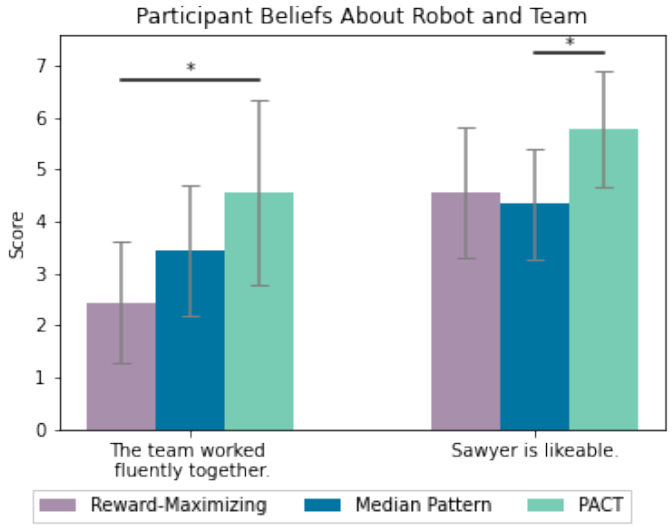

Figure 3.8: Using PACT led to significant improvement in team fluency over baseline ($p = 0.0178$), as well as perceptions of robot likeability over the Median group ($p < 0.05$).

Normalized Scores

Round	Group	Mean Score	p-value
1	Reward-Maximizing	6.81	—
1	Median Pattern	52.00	0.0007
1	**PACT**	**71.27**	**0.0**
2	Reward-Maximizing	27.12	—
2	Median Pattern	60.33	0.0246
2	**PACT**	**78.68**	**0.0006**
3	Reward-Maximizing	21.74	—
3	Median Pattern	79.49	0.0012
3	**PACT**	**88.06**	**0.0003**

Table 3.1: Normalized game scores and p-values obtained via Tukey's HSD for each pattern-based group compared to the baseline Reward-Maximizing group. There were no significant differences between the PACT and Median groups for normalized scores across all rounds.

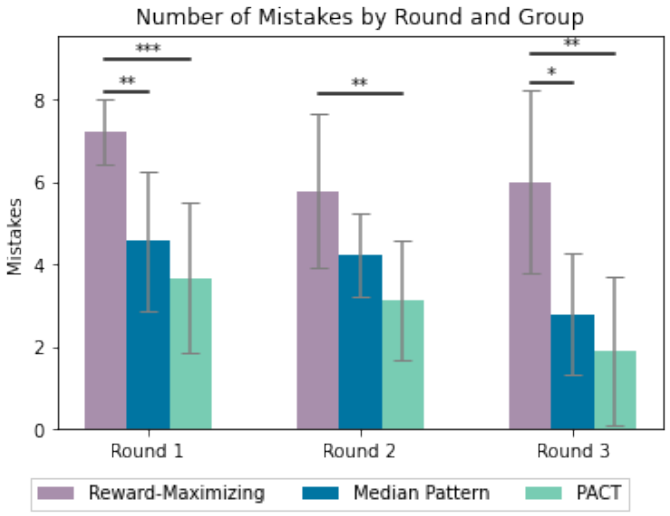

Figure 3.9: Participants in the group that engaged with a robot using PACT made significantly fewer mistakes than the baseline group across all three rounds ($p = 0.0003, 0.0042, 0.0005$), whereas the Median group only made significantly fewer mistakes in two rounds ($p = 0.0047, 0.1138, 0.0053$).

by Table 3.1. As seen in Figure 3.9, the PACT Pattern group made significantly fewer errors when compared to the Reward-Maximizing group across all rounds of gameplay. The Median Pattern group made significantly fewer errors than the Reward-Maximizing group in rounds one and three, but there was no significant difference over the Reward-Maximizing in round two. Part of this may be due to the differences in the patterns seen by each group. Participants in the Median Pattern group saw much more ambiguous patterns than those in the PACT group, meaning that participants in the Median Pattern group could play at least half of the first round and obtain a perfect score by following a pattern other than the robot's pattern. The Median Pattern group was the only group to show significance over the Reward-Maximizing after only one round of gameplay in their belief of understanding how the robot was choosing blocks (p=0.0241), but this effect was no longer significant after another round of gameplay.

Further validating $\mathbf{H_3}$, participants rated the predictability and understandability of the robot in a variety of questions in the Post-Experiment Survey. Using Tukey's HSD, comparisons between the Reward-Maximizing group and both patterns-based groups were significant (Figure 3.10). When compared to the Reward-Maximizing baseline, participants in both the PACT Pattern group ($p < 0.0001$) and the Median Pattern group ($p = 0.0001$) felt the robot was predictable. When asked about the understandability of the robot's actions, the PACT Pattern group ($p = 0.0003$) and the Median Pattern group ($p = 0.0097$) both felt the robot was understandable compared to the Reward-Maximizing baseline. However, there is an important caveat to this finding. Participants were asked about the broader application of the system, and whether they believed most people would be able to understand the robot (Figure 3.10). **Only participants in the PACT group felt that most people would be able to understand the robot**, compared to both the Reward-Maximizing group ($p = 0.0009$) and Median Pattern group ($p = 0.0441$), confirming the premise of this work and validating the proposed contribution. While participants in the Median Pattern group believed at the end of gameplay that they understood the robot's decisions, they did not believe that the system they saw would be broadly understandable.

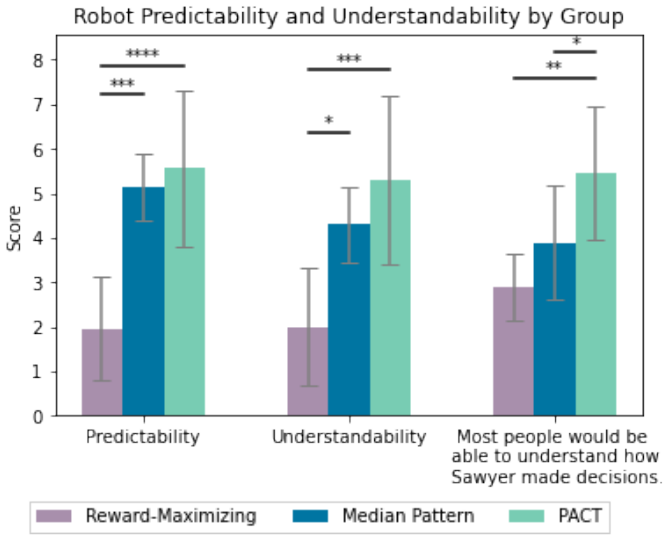

Figure 3.10: Participants in the PACT Pattern ($p < 0.0001$) and the Median Pattern ($p = 0.0001$) both found the robot significantly more predictable than the baseline. Both groups also found the robot's behavior more understandable than the baseline group. (PACT $p = 0.0003$, Median $p = 0.0097$) Only the participants who used PACT felt the robot would be broadly understandable to people when compared to the baseline ($p = 0.0009$) as well as the Median Pattern group ($p = 0.0441$).

3.5.1 Discussion

Our results support the claim that PACT allows a robot to schedule its tasks more predictably, allowing humans to work more effectively with it. This effectiveness stems from the deep-seated human tendency towards pattern recognition and usage. Evidence of this unconscious tendency emerged in participant exit interviews. Despite the lack of a human-visible pattern in the Reward-Maximizing group, approximately half of the participants were convinced that the robot was engaging in some pattern or rule-based behavior. Many voiced that given more gameplay, they likely would be able to find the pattern in the robot's behavior. Many of these participants indicated that they were searching for a pattern that "must" be there, despite there not being any observable pattern.

Additionally, the majority of participants who saw a pattern were unable to articulate the pattern or to fully explain the robot's behavior. Even in the group that saw the PACT pattern, less than half of participants could fully explain the pattern they saw, despite many of them playing perfectly coordinated rounds with the robot.

Anecdotally, this may indicate that centering human cognition and reasoning leads to more unconscious decision-making by humans. Perhaps participants who see a PACT pattern are able to unconsciously predict the robot's next move, without having to use logic or more complex reasoning. Further work to explore this phenomenon and its impacts on human-robot teaming is necessary.

3.6 Conclusions

Participants who collaborate with a robot whose behavior follows pattern-based behavioral conventions selected via PACT report significantly better subjective (perceptions of the robot) and objective (scores) measures when compared to participants who collaborate with a robot focused solely on maximizing team reward. Participants who engage with a robot that uses pattern-based behavioral conventions that are not optimized for the environment by PACT still realize significant performance improvement in coordination, but at the expense of subjective perceptions of the

collaboration and robot. These study results reinforce the importance of leveraging convention in fluent human-robot collaboration, and confirm that PACT is an effective mechanism to do so.

This work demonstrates that intentionally leaning into human cognitive tendencies and de-emphasizing reward-maximizing behavior leads to substantially better outcomes along both objective and subjective metrics. Our proposed method does not preclude the usage of other planning tools, and can be used in tandem with other methods to make robots more predictable while remaining capable. Additionally, the tradeoff between optimal planning and predictability can be negotiated for any environment; PACT can create complex patterns similar to optimal planning, or simple ones to maximize predictability.

As robots are placed in environments where they will be trusted with a diversity of tasks, especially in cases where they will be in close contact with humans, it is critical to characterize and address the disparities between the way robots and humans reason. Robots that are exclusively optimizing for a given reward are reasoning about their environment and collaborations in a fundamentally different way from the humans that they work around and attempt to collaborate with. This leads to a lack of predictability, limiting collaboration. **PACT demonstrates that we can bridge this gap and make robots more predictable without limiting team performance.**

3.7 Transitioning to Continuous Spaces and Balancing with Optimality

This work demonstrates that human-observable patterns can be meaningfully codified for robotic systems and incorporated into robot task-level planning. By explicitly representing patterns as a stacking of rules and ordered features, robots can make decisions that are more predictable and understandable during interaction. The results indicate that patterns do, in fact, matter: incorporating them leads to meaningful differences in human perceptions of robots, their predictability, and their views on a robot-human team they are part of. In particular, this work, via PACT shows promise as a method for selecting effective patterns, indicating that robots can not only use patterns, but also choose among them in a principled way, as the study indicates that simply using a human-legible pattern does not lead to the full spectrum of benefits that selecting

an ideal pattern via PACT provides.

However, the current approach leaves several questions unexamined. It is constrained to discrete task-level planning, which does not fully capture the continuous and dynamic nature of real-world human-robot interaction. A more complete solution should extend these ideas into continuous planning and control, where patterns can influence behavior at a finer temporal scale. Additionally, the current comparison is primarily between pattern use and no pattern use at all, which is an important first step but does not address the more nuanced question of how to balance patterning and optimality. The subsequent work in this thesis seeks to strike a balance between predictability and optimality by minimizing a combined cost that quantifies optimality and patterning within a continuous domain.

Chapter 4

Thinking in Patterns: Sacrificing Performance for Predictability Enhances Human-AI Teams

After confirming the viability of utilizing patterns in human-robot teaming, we need to extend pattern-based planning beyond discrete task selection and into continuous decision-making and control. Real-world robot behavior unfolds in a continuous space of motion, timing, and interaction, so enabling robots to plan with patterns at this level is essential for producing behavior that is more broadly usable. This also requires developing a representation of patterns that is suitable for continuous spaces; capturing not just high-level tasks, but also full trajectories, spatial relationships, and temporal dynamics in a way that can guide patterned robot behavior in continuous space.

At the same time, we must also address the important tradeoff between optimality and adherence to recognizable patterns. As previously discussed,strictly optimizing for efficiency or task completion often leads to behavior that is difficult for people to anticipate, while solely following patterns has an impact on the robot's ability to maximize reward. The key challenge is therefore to strike an effective balance: designing methods that provide the benefits of predictable, pattern-consistent behavior while still preserving the efficiency and high reward of SOTA methods.

4.1 Introduction

Human-AI teaming is hindered by an inescapable conflict between the logic of machine optimization and the foundations of human cognition. While algorithms can generate highly efficient behaviors, these actions often lack a structure rooted in the same cognitive toolkit of patterns and

heuristics that guides human reasoning [114, 40, 78] that people rely on to build predictive models of their partners [46, 81, 34, 77, 83]. This disconnect between the robot's complex, optimal trajectory and a human's heuristics-based expectations creates a 'predictability gap', which in turn compromises team fluency and erodes collaboration [71]. *This core design tension is not unique to robotics* but, rather, a universal challenge in human-autonomy interaction: across autonomy, machine-optimal actions often read as unpredictable to people. Autonomous vehicles that brake or merge on mathematically efficient profiles can feel abrupt and inscrutable to drivers and pedestrians, eroding trust and coordination [62, 33, 82]. Clinical and financial decision-support systems may output performance-optimal recommendations that experts cannot anticipate or explain, inhibiting adoption despite their accuracy [60, 75, 123]. *The core problem is not that autonomous systems are different, but that autonomy 'reasons' differently than people do* [114, 115].

In this work, we posit that bridging this divide requires more than just algorithmic improvements; it requires aligning autonomous behavior with the foundational mechanisms of human understanding. This informs the formulation of a clear, testable hypothesis: if human teaming is built upon a cognitive toolkit of pattern recognition, then intentionally embedding simple, perceivable structure into a robot's actions should significantly improve predictability and team fluency, even if regulated by a bounded loss of individual task performance on the robot's part. To investigate this, *patterns must be treated not as byproducts of simplification, but as a design variable to be precisely controlled.*

This requires a new scientific instrument capable of systematically tuning the level of perceivable structure in an agent's behavior while bounding performance loss. To this end, we introduce Predictability-Satisfying Trajectory Optimization (PRESTO), a human-aligned trajectory optimization adaptation layer that introduces patterns into behaviors generated by other controllers. This work leverages PRESTO as a scientific instrument to characterize the predictability-performance tradeoff, revealing the surprising result that a (bounded) performance sacrifice in service of predictability has an outsized benefit for team performance metrics. We provide **the first direct empirical measurement of this tradeoff's effects on team fluency**, specifically

in the context of adapting reinforcement learning policies. Across two human-subjects studies, our findings reveal fundamental principles for human-autonomy interaction. First, we demonstrate a **decoupling of perceived intelligence from behavioral complexity**, challenging the assumption that mathematically optimal or complex actions are perceived as the most intelligent [118, 113, 7, 53, 86]. Second, by tracking participants' predictions over time, we offer novel, quantitative evidence that patterned behavior fosters **more stable and accurate human mental models** of an autonomous partner, a cornerstone of effective team cognition. Together, these results establish that predictability, not algorithmic complexity, is the key driver of fluent human-autonomy teaming.

4.1.1 Technical Content and Approach

To resolve the conflict between machine optimization and human cognition, our work synthesizes a convergence of findings from cognitive science and Human-Robot Interaction (HRI) to articulate a central design principle: *designing for predictability*. This principle holds that predictability is one of the main mechanisms that allows a human to form an accurate *mental model* of their robotic partner, a cognitive structure representing its intentions and future actions [31, 108, 11]. An accurate mental model is a cornerstone of effective team cognition [79, 90, 56], and the HRI literature has consistently established that predictable robot behavior enhances team fluency, efficiency, and trust [27, 71, 65, 36]. Therefore, designing for predictability is not merely a user preference; *it is a direct mechanism for improving the cognitive alignment between human and machine*.

Yet, this principle runs directly counter to the foundational logic of the very tools used to create high-performance robot behavior. Methods like reinforcement learning (RL) excel at searching for and discovering optimal, reward-maximizing policies and have been widely applied in human-agent teaming [107, 39, 106, 51, 5, 4]. The fundamental issue is that the reward functions guiding these policies are typically agent-centric, evaluating the immediate outcome of an action without considering how that action is perceived by an external observer as part of a larger temporal

sequence. Even if one could engineer a reward function to account for an external observer, a more foundational limitation remains: the reliance on the Markov assumption. This assumption makes it very expensive for RL agents to reason over their trajectories; they optimize for the next best action from the current state, making them unaware of the history of their own actions (without otherwise introducing these data into their state space). Thus, standard RL objectives may optimize for task reward without being able to control observer predictability or repeated temporal structure as a design variable. *This produces behaviors that, while locally optimal, are often complex and deficient in the patterns humans need to form a stable predictive model.*

Addressing this issue requires a fundamentally different design philosophy: one that shapes the **global structure of behavior**, rather than optimizing a sequence of local decisions. This perspective is crucial because human partners do not perceive robot actions in isolation; they interpret the entire motion sequence to infer intent and predict future behavior, an interpretation heavily shaped by a history of prior interactions and innate cognitive biases. In this work, the technical embodiment of this philosophy is trajectory optimization: a method that treats an entire trajectory as the unit of design. *Its ability to reason about a sequence as a whole positions it as the appropriate tool for sequence-level regularization to encode the very temporal qualities that Markovian methods are ill suited for*, such as expressing high-level intent, adhering to social conventions, and improving predictability [37, 64, 54]. This approach has proven effective across a range of applications for generating behavior that aligns with human expectations [92, 41, 96].

Resolving this dichotomy requires a principled synthesis: an approach that retains the power of optimal policies while layering on the global, human-centric structure they lack. Our work operationalizes this principle through PRESTO (PREdictability-Satisfying Trajectory Optimization): as an optimization layer, PRESTO takes an existing behavior generator (e.g., a policy created by an RL algorithm) and injects human-perceptible patterns into this behavior via trajectory optimization over rollouts of the policy. By explicitly balancing a *predictability* cost against a *deviation* cost under a bounded performance loss, it provides the precise controls needed to navigate the Pareto front between predictability and performance. *This allows us to treat patterning as a core design*

variable, enabling the direct, empirical test of our central hypothesis: that aligning autonomy with human cognition, not raw algorithmic complexity, is the key to unlocking fluent collaboration.

4.2 Results

4.2.1 Experiment Overview

To investigate how aligning autonomous behavior with human cognitive patterns affects team fluency, we conducted two between-subjects experiments. Both studies were approved by an Institutional Review Board (IRB). The first study investigated this phenomena through an on-screen map (online study; bird's eye view) similar to waypoint-based interaction paradigms for remotely operating robots across large length-scales. The second study investigated the phenomena through an immersive virtual reality (VR) outdoor environment (in-person study; first-person view). Participants were randomly assigned to collaborate with an agent exhibiting one of two distinct behavioral philosophies. In the **optimal** condition, the agent executed a reward-maximizing path generated by a state-of-the-art reinforcement learning policy, representing pure machine-centric efficiency. In the **patterned** condition, this same optimal policy was adapted to generate behavior that conforms to simple, human-perceptible patterns, representing a human-centric approach.

4.2.1.1 Design

In the online experiment, 120 participants were recruited through the Prolific online experiment platform and were randomly assigned to two conditions: optimal and patterned. Each participant completed ten rounds of a task in which they attempted to predict a robot's path over an on-screen bird's eye view of a map, with the agent's trajectories generated either by the optimal reinforcement learning policy directly, or by adapting that policy's output to conform to human-perceptible patterns. In each of the ten rounds, participants repeatedly predicted the agent's path to a goal. After each prediction, a segment of the agent's true path was revealed, and participants updated their prediction for the remainder of the trajectory (Figure 4.2).

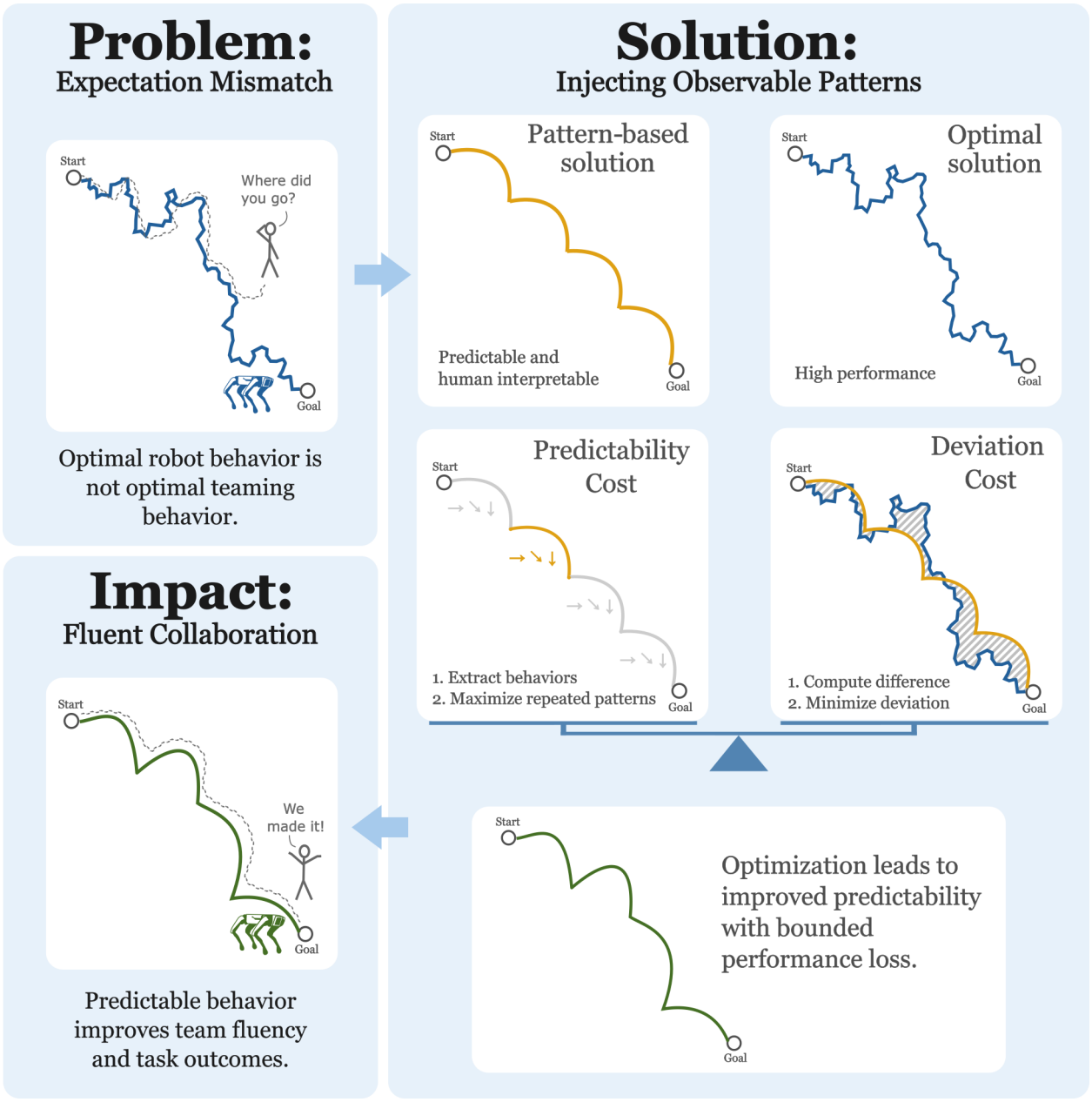

Figure 4.1: **An overview of pattern-driven behavior.** Pattern-infused behavior bridges the gap between machine optimality and human predictability. Standard reward-maximizing policies (left) often produce behaviors that, while performant, create an expectation mismatch for human partners. Patterned behavior (right) addresses this by injecting observable, repetitive patterns into the robot's motion. By optimizing the trade-off between a predictability cost that maximizes repeated behavioral structures and a deviation cost that minimizes divergence from an optimal policy rollout, we find improved objective and subjective measures of predictability and team performance with only a bounded loss in task performance.

Representative of realistic operating conditions, an information asymmetry existed between the robot and the human. The human is only privy to an approximation of the operating environment as exposed through the birds-eye map, in contrast to the situated robot engaging with the actual environment. The robot's policy considered areas within the environment as either impassable (obstacles), passable with varying negative reward which would slow the robot ('challenging' terrain), and passable neutral areas ('easy' terrain). In the optimal condition, the robot would generate paths maximizing use of easy terrain while prioritizing length efficiency. Paths in the patterned condition, however, could cross into challenging terrain if doing so would make the path more predictable. The information asymmetry arises from the fact that participants were only able to visually distinguish impassable (obstacles) from passable terrain, with all other areas being generally colored as random shades of green and tan.

In the in-situ VR experiment, 24 participants were recruited from our university community to engage in a teaming activity with a robot in VR. The activity is outlined in Figure 4.5, which shows still images from participant video. Participants were tasked with retrieving a series of rock samples. After collecting each sample, they had to intercept their continuously moving robotic partner to deposit it—a task that required them to implicitly predict the agent's future location for a successful rendezvous. Upon completing a retrieval, the next sample location would be visualized to participants, and the process repeated. Participants retrieved five samples in each round of gameplay and participated in five rounds of the activity in total. Each round occurred within a different environment.

As in the online experiment, robots in the optimal condition avoided challenging terrain wherever possible, per the optimization criteria of the policy used. However, in this experiment, these areas were visually distinguishable from surrounding areas. Challenging terrain (areas of negative reward) were represented via the use of taller grass, sandy terrain, and rocky areas. In this way, the information asymmetry was reduced, as participants could observe the robot avoiding such areas or slowing down when it crossed them. Of note, however, is the fact that this aspect of the robot's policy was not explained to participants explicitly, and the resolution of this information

asymmetry relied upon the participants' ability to observe the robot and accurately model its behavior.

4.2.1.2 Hypotheses

The two studies test hypotheses centered on three core psychological and performance outcomes: the accuracy of human prediction, the cognitive load experienced by the human partner, and the subjective perception of the autonomous agent.

In the online study, in which participants had a bird's eye view of the agent and explicitly predicted the robot's path, hypotheses cover both the self-reported cognitive fatigue and perceptions of the robot as well as objective metrics of prediction accuracy.

- H_1: Participants will be more accurate in predicting patterned trajectories than task-optimal trajectories.
- H_2: Participants will self-report lower scores for cognitive fatigue when predicting patterned trajectories than task-optimal trajectories.
- H_3: Participants who interact with patterned trajectories will view the agent as both more predictable and more understandable than those who interact with task-optimal trajectories.

In the in-situ environment, participants implicitly predicted the robot's behavior by anticipating its future location in deciding how to return to the robot.

- H_4: Participants who team with a robot using patterned trajectories will be able to navigate back to robot more quickly after separation than those who team with a strictly task-optimal robot.
- H_5: Participants who work with robots using patterned trajectories will report lower scores for cognitive fatigue than participants in the task-optimal condition.

- H_6: Participants who engage with the patterned trajectories will view the robot as more predictable and understandable than participants in the task-optimal group.

4.2.2 Patterned motion enables accurate, low-effort prediction from an observer's perspective (H_1, H_2, H_3)

120 participants completed an online study where they directly predicted the robot's path to a known goal. Data from four participants were excluded due to missing information from multiple rounds.

4.2.2.1 Patterned behavior aligns with human intuition, improving prediction accuracy (H_1):

Aligning the robot's trajectory with human-centric patterns had a significant effect on prediction accuracy, validating H1. Post-hoc comparisons indicate that participants in the pattern condition predicted robot trajectories that were closer to the ground truth of the robot's path, regardless of round or proportion of the trajectory seen, shown in Figure 4.3. There is no learning effect present for participants in the pattern condition. Their high initial accuracy left little room for improvement ($p = 0.834, M_1 = 57.84, M_2 = 53.66, d = 0.08$). In contrast, participants in the optimal condition became significantly less accurate over time ($p = 0.023, M_1 = 67.78, M_2 = 77.89, d = 0.16$). These data suggest that patterned trajectories aligned with innate human expectations from the outset. This performance degradation in the optimal condition is consistent with prior work on predictability, where ambiguity or a lack of understandability in the robot's behavior may lead to an incorrect mental model of the robot and a commensurate decrease in performance over time [71].

4.2.2.2 Predictable motion reduces the cognitive cost of collaboration (H_2)

Observing patterned, predictable motion significantly reduced the cognitive fatigue associated with the prediction task, validating H2 (Figure 4.3). Participants in the

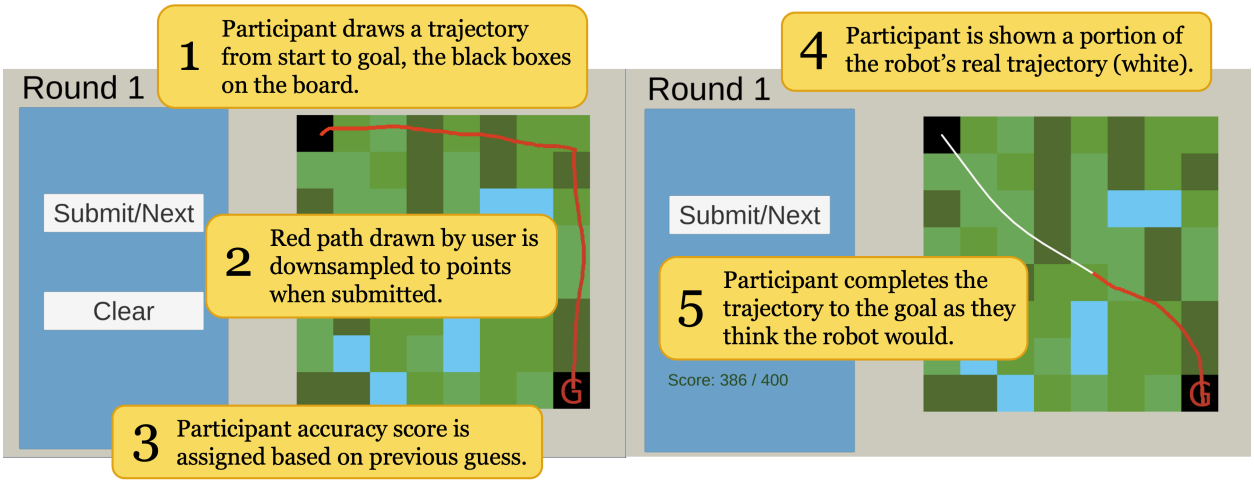

Figure 4.2: **Annotated screenshots from the user interface of the online study.** Participants must predict a robot's path by drawing on a map. Each round consists of a sequence of trials that progressively reveal more of the robot's actual path from start to goal (white) in 16.67% increments, beginning with none of the path revealed. (Left) Participants draw a path prediction predicting the completion of the path (red). (Right) Participants were scored based on the accuracy of their path predictions to the ground truth path actually taken by the robot.

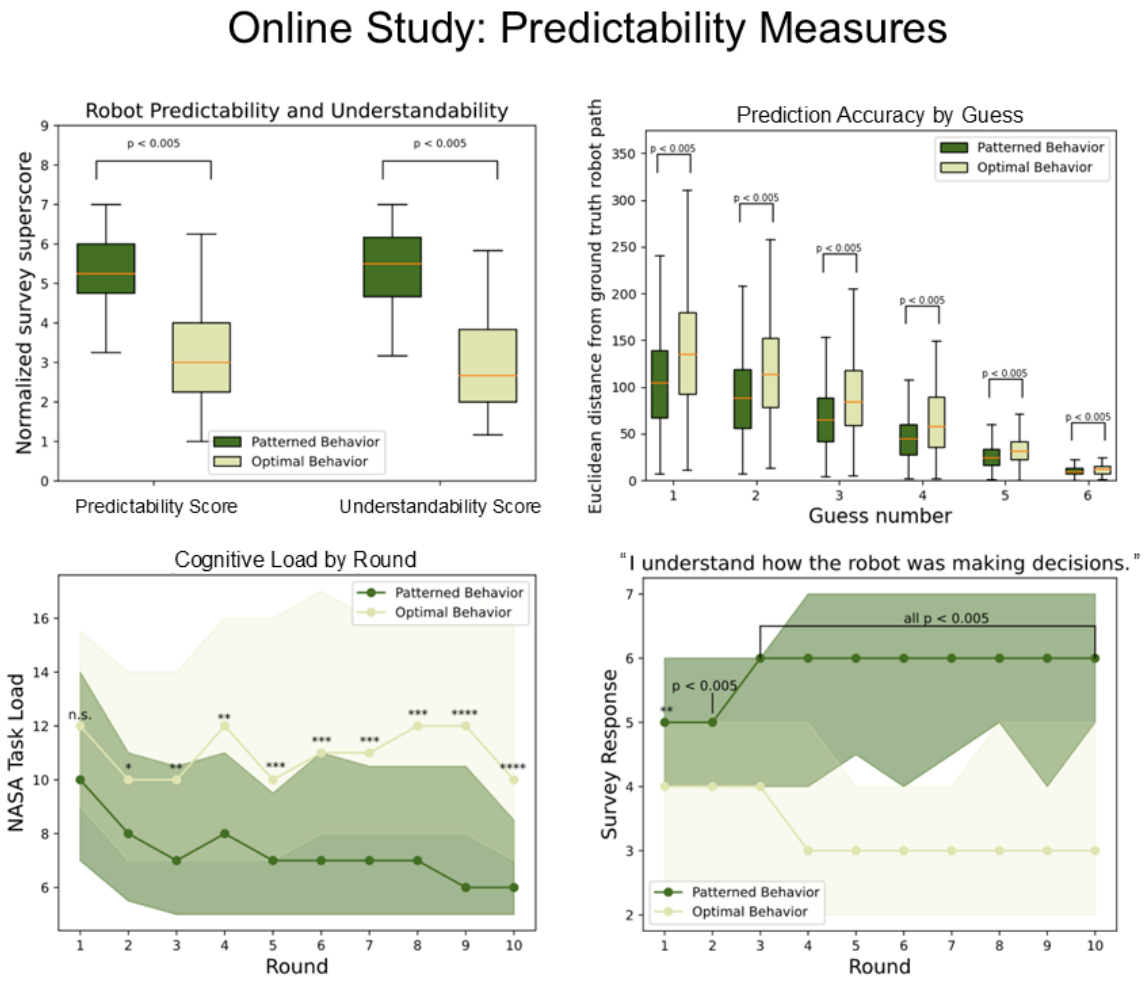

Figure 4.3: **Agents exhibiting patterned motion were perceived as significantly more predictable according to both quantitative and qualitative measures.** Participants in the patterned condition were significantly more accurate at predicting the remainder of the trajectory in every prediction. These participants self-reported significantly higher scores for the robot being predictable ($p < 0.005, M_1 = 5.33, M_2 = 2.99, d = 2.19$) and understandable ($p < 0.005, M_1 = 5.34, M_2 = 2.89, d = 2.14$) than participants did in the optimal motion condition. Participants also self-reported lower cognitive fatigue scores when interacting with a robot using patterned motion ($p < 0.005, M_1 = 8.58, M_2 = 11.72, d = 0.66$). When prompted between rounds for their understanding of the robot's decision making process, participants who saw patterned paths rated their understanding as significantly higher than those in the optimal condition for all rounds ($p < 0.005, M_1 = 5.33, M_2 = 3.40, d = 1.18$).

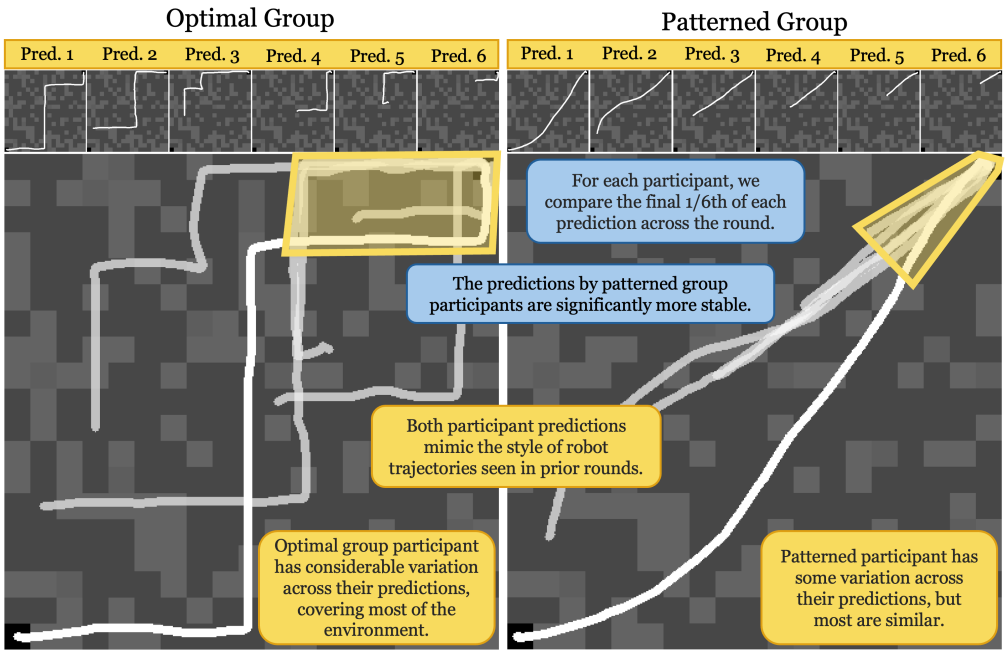

Figure 4.4: **Repeatedly asking participants to predict the same trajectory with increasing amounts of information uncovered a novel effect of human-aligned trajectory optimization: participants in the patterned condition had more stable predictions.** Participants were asked to predict the robot's trajectory six times in the same environment, each time with more information. Participants in both conditions adjusted their predictions of the robot's behavior throughout the round, but the patterned condition participants were significantly more stable in their perception of the robot's trajectory ($p < 0.005, M_1 = 25.04, M_2 = 38.43, d = 0.31$) — meaning their predictions changed much less. As shown in Figure 4.3, participants seeing patterned behavior subjectively indicated higher confidence in their understanding of the robot's decision making process, quantitatively affirmed here through their behavior. As their initial intuition about how the robot would approach the goal was correct, more information reinforced their beliefs rather than challenged them.

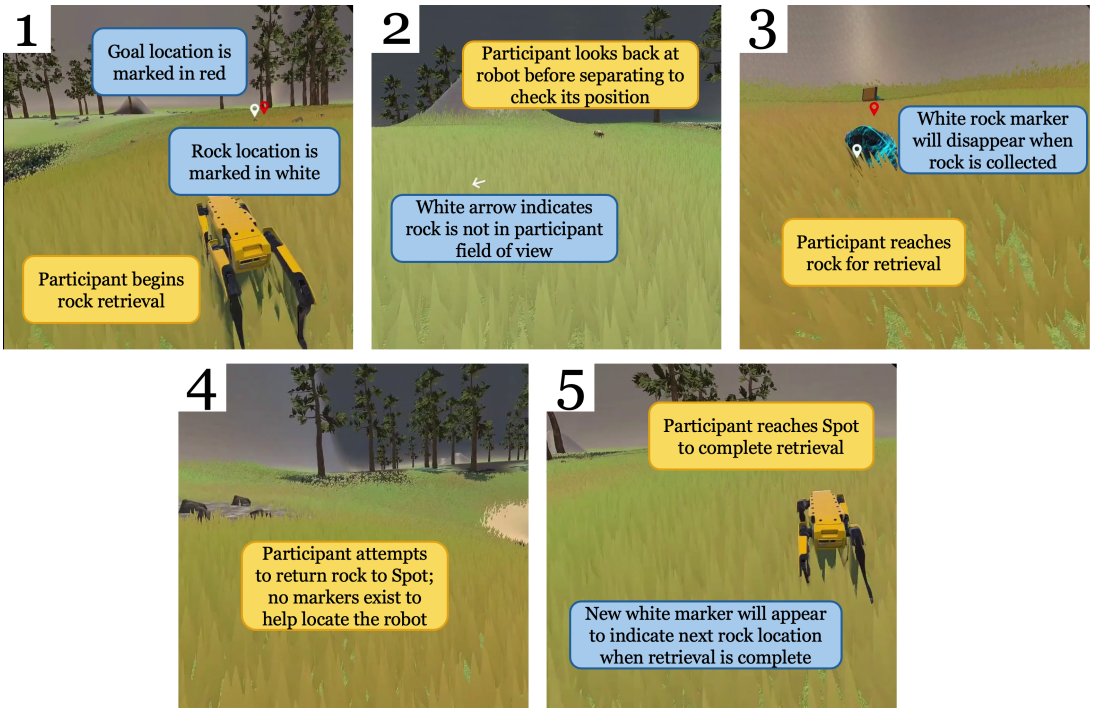

Figure 4.5: **A step-by-step example of the rock retrieval task in the virtual reality simulation.** At the start of the round, two markers are visible to the participant. A red marker indicating the goal location is always present. A white marker indicates the location of the sample that the Spot robot has flagged for retrieval. When the participant departs from Spot and begins the retrieval (2), they may look back to check on the robot's location. When the sample is outside of the participant's field-of-view, an arrow appears indicating where the participant should turn to find it. In panel 3, the sample is visible along with the robot's goal. Once the participant retrieves the sample, the marker disappears (4) and they must return to the robot. The robot location is not indicated by a marker; participants must predict where the robot is and find it within the environment (5).

patterned condition reported a significantly lower level of cognitive fatigue at the end of round ten compared to their first experience predicting the robot's trajectory ($p < 0.001, M_1 = 10.82, M_2 = 7.64, d = 0.74$), whereas the optimal condition does not ($p = 0.369, M_1 = 12.35, M_2 = 11.75, d = 0.11$). This finding suggests that by conforming to simple patterns, the robot's behavior engaged more intuitive and less effortful human cognitive processes, such as pattern completion, rather than requiring demanding analytical reasoning.

4.2.2.3 Patterned motion creates shared, stable mental models and drives positive team perception (H_3)

Participants who observed patterned agents formed a more predictable and understandable model of their behavior, validating H3. In line with prior pattern-driven predictability work [71], participants in the patterned condition made decisions significantly faster than optimal condition participants ($p < 0.001, M_1 = 8149.12, M_2 = 10651.49, d = 0.11$). Participants additionally perceived the robot as more likeable ($p < 0.001, M_1 = 18.76, M_2 = 15.25, d = 0.79$) according to the RoSAS likeability subscale [16]. These results illustrate the strong effect of increased predictability on participants' perceptions of the human-robot team. Participants who observed these human-aligned, patterned trajectories had significantly more positive perceptions of the robot as an effective teammate ($p < 0.001, M_1 = 5.14, M_2 = 3.86, d = 1.05$), as well as of the team's fluency ($p < 0.001, M_1 = 5.07, M_2 = 3.56, d = 1.03$) and coordination ($p < 0.001, M_1 = 5.20, M_2 = 3.84, d = 0.90$). These findings further strengthen the connection between predictability and effective human-robot teaming.

A key scientific insight from this study is the decoupling of perceived intelligence from algorithmic complexity. Participants judged the robot exhibiting patterned behavior as significantly more intelligent ($p < 0.001, M_1 = 21.9, M_2 = 17.3, d = 1.07$) than the one executing mathematically optimal paths. This finding challenges the common assumption that behavioral complexity is a proxy for intelligence in artificial agents, suggesting instead that in collaborative contexts, predictability is a primary marker of effective intelligence.

Study participants made several consecutive predictions about the same trajectory — providing multiple completions of the later portions of the robot's path. The changes between these guesses were analyzed for both groups, shown in Figure 4.4. While both groups altered their predictions of the robot's behavior when given more information, the pattern condition participants modified their predictions significantly less than those in the optimal condition ($p < 0.001, M_1 = 25.04, M_2 = 38.43, d = 0.31$). This result provides direct, quantitative evidence for the formation of a stable mental model, a cornerstone of effective team cognition [116]. Patterned trajectories enabled participants to form robust initial predictions that were reinforced, rather than contradicted, by subsequent information. This demonstrates that algorithmic choices can directly support a human's cognitive process of building and maintaining an accurate model of their autonomous partner.

Participants were asked multiple related questions about their perceptions of the predictability and understandability of the robot. To analyze these questions, we create composite scores for predictability and understandability. For the predictability composite score, four related questions (see Appendix) are aggregated. A Cronbach's alpha of $\alpha = 0.89$ indicates high internal consistency, supporting the validity of combining these items into a single scale. Similarly, six questions (listed in Appendix) are summed for the understandability composite score ($\alpha = 0.94$). Both composite scores are scaled by the number of questions used to form the sum, the results of which are shown in Figure 4.3. Participants in the patterned condition found the agent significantly more predictable and its behavior more broadly understandable than those in the optimal condition did ($p < 0.005, M_1 = 5.33, M_2 = 2.99, d = 2.19, p < 0.005, M_1 = 5.34, M_2 = 2.89, d = 2.14$). **These results validate H3. Participants' subjective ratings confirm the patterned agent was perceived as significantly more predictable and understandable, and this perceived understanding was objectively substantiated by their behavior: they formed more stable, robust mental models that were reinforced, not contradicted, by new information.**

4.2.3 Predictable behavior enables fluent physical collaboration in an embodied task (H_4, H_5, H_6)

In the in-person VR experiment, data was collected from 24 participants as part of an IRB-approved study. Two participants withdrew early due to nausea caused by VR, a proportion lower than documented rates of cybersickness [22]. Though these participants did not complete all five rounds, their data for the rounds they did complete as well as their post-experiment survey and semi-structured interview data was used in the analysis, as all data collected was within norms for their experimental group. Screening between rounds prevented participants from continuing at the first signs of cybersickness, preventing a drop in performance due to illness.

4.2.3.1 Shared understanding of motion patterns improves team performance (H_4)

The shared understanding enabled by patterned motion translated directly into superior team performance (Figure 4.7), validating H_4. Participants who worked with a robot in the patterned condition were able to complete their retrieval tasks significantly faster ($p < 0.001, M_1 = 14.91, M_2 = 21.43, d = 0.37$), while traveling less distance in the environment ($p = 0.0037, M_1 = 55.16, M_2 = 77.49, d = 0.33$), and along a more direct path ($p < 0.001, M_1 = 13.84, M_2 = 29.19, d = 0.25$) (measured by the deviation from the straight-line path that would intercept the robot) than those in the optimal condition. Other than the first retrieval, which was designed to be simple, participants in the pattern condition were significantly faster at navigating back to the robot from the sample location for the last four retrievals ($p_2 < 0.001, M_{1,2} = 13.18, M_{2,2} = 22.45, d_2 = 0.56, p_3 = 0.009, M_{1,3} = 18.46, M_{2,3} = 26.33, d_3 = 0.31, p_4 = 0.002, M_{1,4} = 16.0, M_{2,4} = 26.46, d_4 = 0.62, p_5 = 0.025, M_{1,5} = 15.88, M_{2,5} = 21.66, d_5 = 0.46$). Before retrieving the sample, many optimal participants observed the robot's movements, or turned around to check on the robot. Coded video recordings of participants were used to measure the amount of time spent checking on the robot. Participants in the pattern condition spent significantly less time checking on the robot per round ($p = 0.004, M_1 = 5.39, M_2 = 12.75, d = 0.399$). They also spent less time

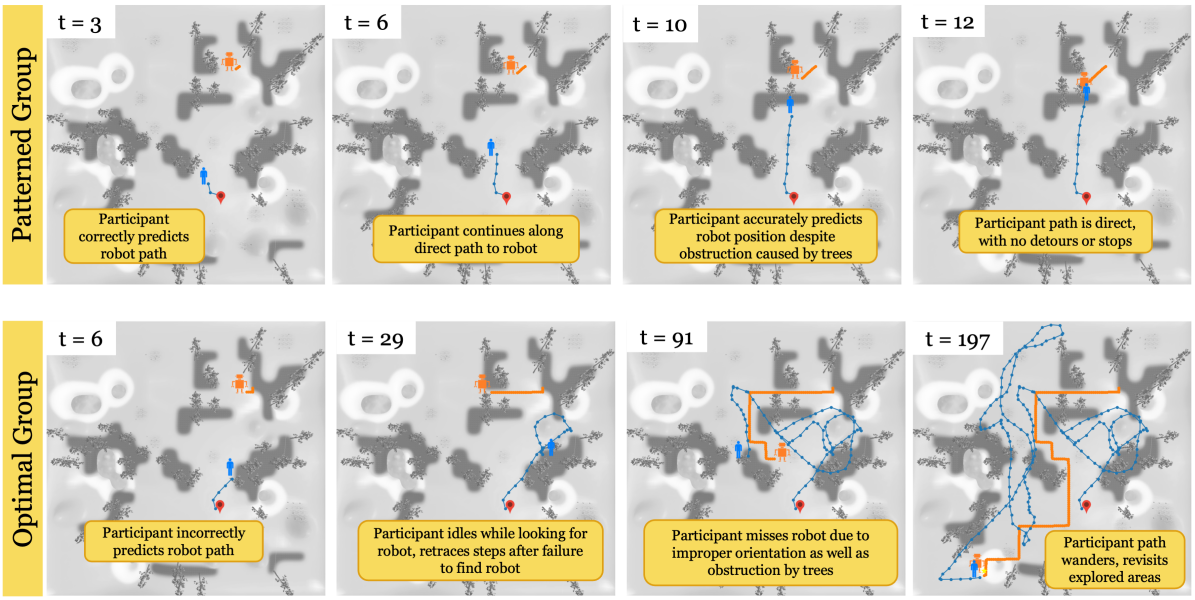

Figure 4.6: **Participants in the patterned condition traveled along more direct paths when implicitly predicting the robot's location.** This figure compares two different robot behaviors in the same environment. In both cases, participants begin in the same place and the robot has a similar starting location. The initial prediction of the patterned group participant is correct, but the optimal group participant's is not. As in this example, participants in the patterned condition generally traveled along significantly more direct paths as compared to the optimal group ($p = 0.009, M_1 = 13.84, M_2 = 29.19, d = 0.25$). Here, the optimal group participant makes an incorrect prediction at the start of the rendezvous and struggles to rectify this, taking over three minutes ($t = 197s$) to locate the robot in the virtual environment.

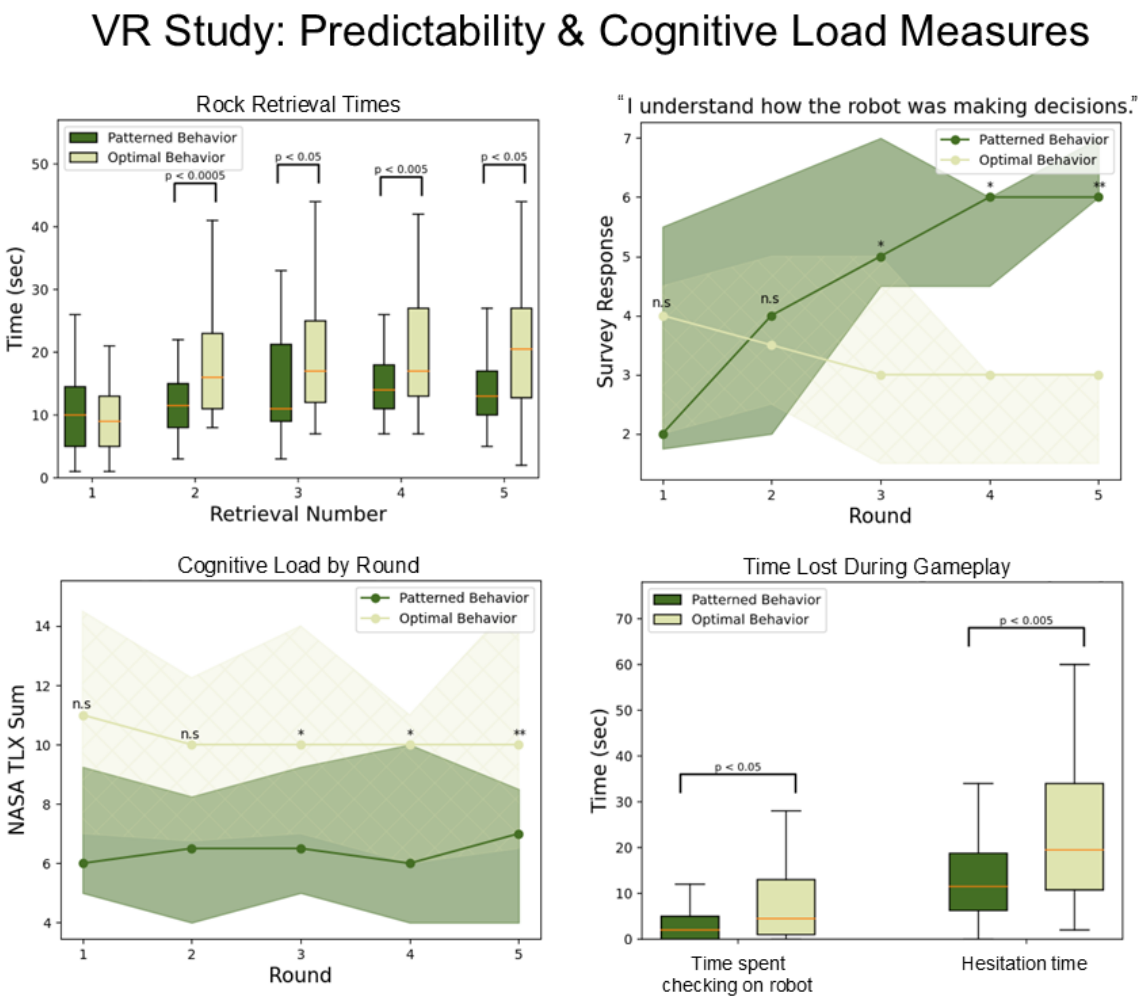

Figure 4.7: **Participants in the pattern condition were better at predicting the robot's behavior, and did so with less cognitive fatigue.** Between rounds of gameplay, participants were surveyed about their cognitive fatigue and understanding of the robot's decision making process. After three rounds of working with the robot in complex environments (approximately 12 minutes), participants in the patterned condition were significantly less fatigued and rated their understanding of the robot significantly higher.

hesitating during retrievals, idling as they decided where to go, and engaged in fewer directional changes along their paths ($p < 0.001, M_1 = 4.24, M_2 = 6.9, d = 0.33$). Furthermore, participants collaborating with a patterned agent demonstrated a strong, task-beneficial learning effect, improving their speed ($p = 0.002, M_1 = 17.223, M_2 = 12.73, d = 0.57$) and efficiency ($p = 0.009, M_1 = 31.22, M_2 = 48.09, d = 0.55$) over time. There is no evidence that such learning occurred in the optimal condition ($p = 0.334, M_1 = 3.42, M_2 = 4.05, p = 0.315, M_1 = 45.02, M_2 = 50.444$). **This divergence indicates that patterned motion facilitates the development of an accurate mental model, a process that failed to occur when participants observed purely optimal behavior.**

4.2.3.2 Predictability reduces attentional demands and collaborative effort (H_5)

Overall, participants in the pattern condition reported significantly lower scores across the NASA TLX (Figure 4.7), indicating significantly less cognitive fatigue from working with the robot compared to those in the optimal behavior condition ($p < 0.001, M_1 = 7.48, M_2 = 10.64, d = 0.744$), validating H_5. Whether participants were interacting through a bird's-eye view or were co-located with the robot, and independent of the amount of trajectory and environment data given, participants in the patterned behavior condition were less cognitively fatigued by working with the robot and when predicting its path. Post-hoc video coding was performed on video recorded from the VR headset during participant gameplay to determine time spent checking on the robot, such as stopping to look back at the robot or walking backwards to keep the robot in their field-of-view. Analysis of this video coding indicates that participants in the optimal condition spent more time checking on the robot or looking back at the robot when they were retrieving rock samples ($p = 0.013, M_1 = 5.39, M_2 = 12.75, d = 0.40$). Participants in the optimal behavior condition ($M = 12.750$s) spent 7.4 seconds more on average than participants in the patterned behavior condition ($M = 5.386$s) checking on the robot's location. The additional time spent checking on the robot was noticeable and often frustrating to participants, with one in the optimal group noting in their interview that "the rock [is] my second priority and the robot becomes my

first priority."

This demonstrates that behavioral unpredictability imposes a direct attentional cost, forcing the human partner to divert cognitive resources from their primary task to the secondary task of monitoring the agent. While participants in the optimal behavior condition generally indicated that they were aware of their mental model's inaccuracy, they still struggled to improve their performance. Participants in the patterned condition were able to rely on more familiar, intuitive cognitive processes to form a more accurate mental model of their robot teammate.

4.2.3.3 Patterned agents are perceived as better, more cooperative teammates (H_6)

Participants in the patterned behavior condition found the robot significantly more predictable and its behavior more understandable, validating H_6. Participants were surveyed on their perceptions of the robot's predictability and understandability. As in the online study, we created composite scores from related survey questions (Cronbach's $\alpha = 0.93, 0.91$). As shown in Figure 4.8, participants in the patterned condition were significantly more positive in their perceptions of the robot's predictability ($p = 0.0003, M_1 = 5.42, M_2 = 2.92, d = 1.748$) and understandability ($p = 0.0026, M_1 = 5.25, M_2 = 3.12, d = 1.49$).

These positive perceptions extended to the robot's role as a teammate. Participants in the patterned condition perceived the robot's behavior as more aligned with the team's goals ($p = 0.003, M_1 = 5.0, M_2 = 2.58, d = 1.39$), and reported stronger agreement that the robot was a team player ($p = 0.0036, M_1 = 4.58, M_2 = 2.5, d = 1.36$), that the team worked fluently together ($p = 0.032, M_1 = 5.42, M_2 = 3.67, d = 1.05$), and that both robot and human were working towards the same goal ($p = 0.0039, M_1 = 6.33, M_2 = 4.42, d = 1.39$) when compared to the optimal behavior condition.

A crucial finding is surfaced from these results: intuitive predictability is more effective than analytical transparency. In the VR study, the logic for the optimal agent's behavior was visually apparent (e.g., avoiding challenging terrain), yet it was not sufficient for human partners to form an

effective predictive model. In contrast, participants who experienced patterned motion successfully formed accurate mental models (evidenced by consistent, superior performance) even if they could not explicitly articulate the underlying rules. This demonstrates that seeing a simple pattern in the actions themselves is more effective for fluent collaboration than seeing the reason for an agent's complex actions. The success of patterned motion in this embodied domain in addition to the bird's eye domain further evidences that this strategy taps into a core, innate cognitive process for model-building.

4.2.4 A Principle for Human-Autonomy Teaming: Predictability Through Patterns

Across two distinct experimental paradigms, one testing explicit prediction from a detached perspective and the other testing implicit prediction in an embodied team task, a consistent principle emerged: aligning an autonomous agent's behavior with human-perceptible patterns dramatically improves the efficacy and fluency of the human-agent team. Participants collaborating with patterning agents made faster, more accurate decisions and consistently perceived the agent as more predictable and understandable. Additionally, the online experiment shows that patterns enable the formation of more stable perceptions of the robot's behavior across a single interaction. Participants in the virtual reality study reported significantly more positive perceptions of the human-robot team and the robot's ability to be an adequate teammate.

4.2.5 Statistical Analysis

Data were first assessed for normality using the Shapiro-Wilk test. For normally distributed dependent variables, between-group differences were analyzed using independent samples t-tests. For non-normally distributed data, such as Likert-scale survey responses anchored to the top or bottom of the scale, a Kruskal-Wallis test was employed as a non-parametric alternative to ANOVA [61]. As the Kruskal-Wallis test is an omnibus test that does not identify which specific groups differ, significant results were followed by a Dunn's post-hoc test with a Bonferroni correction for multiple

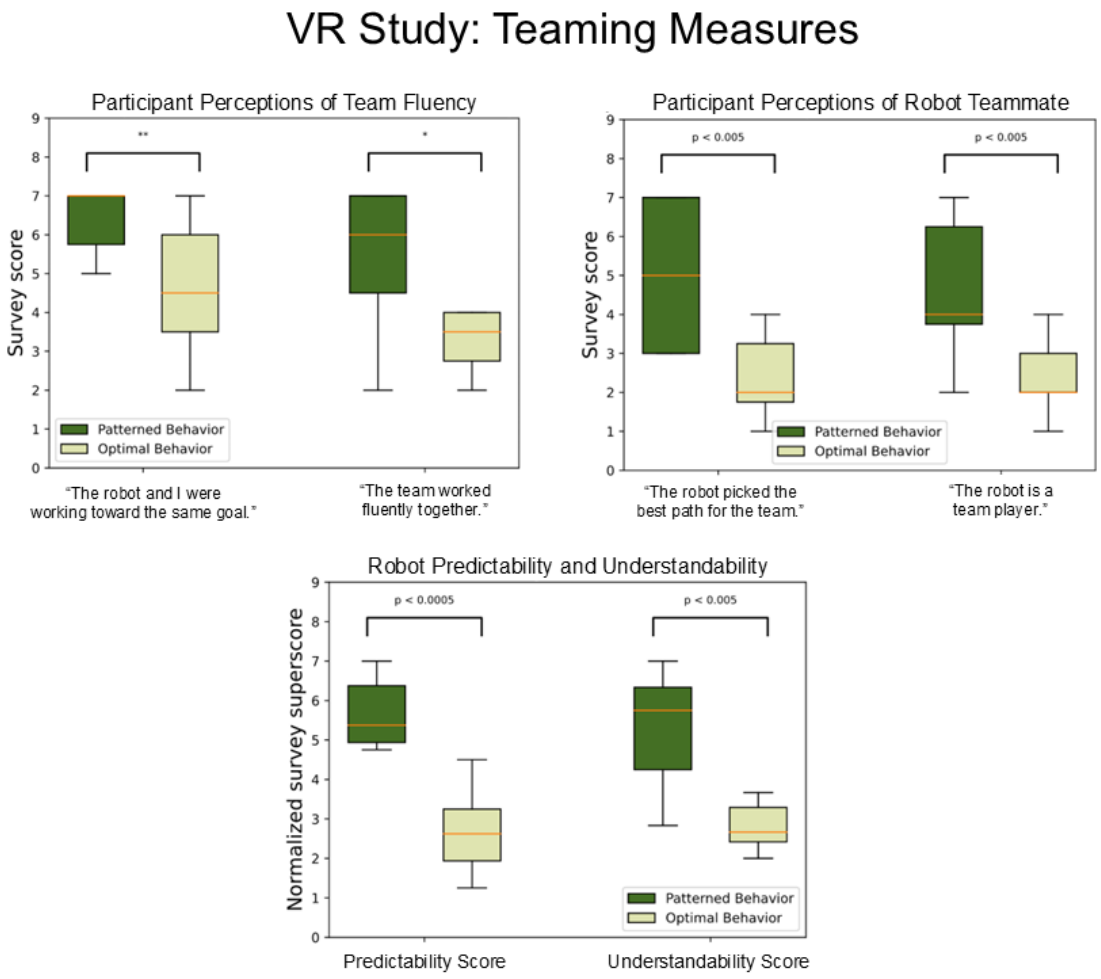

Figure 4.8: **Participants in the patterned condition in the VR experiment found the robot significantly more predictable and understandable than participants in the optimal condition.** Participants who worked with the robot using patterned motion rated the robot significantly more positively as a teammate than participants in the optimal condition, agreeing more strongly that the robot picked the best path for the team and that the robot was a team player. Additionally, participants who interacted with the robot in the pattern condition thought the team worked more fluently together, and agreed that the team was working toward a shared goal.

comparisons to identify significant differences between the patterned and optimal conditions. All statistical tests used an alpha level of 0.05.

4.3 Discussion

Our findings establish a core principle for the design of collaborative autonomous systems: aligning an agent's behavior with innate human cognitive faculties for pattern recognition yields substantial improvements in team fluency, efficiency, and trust, even when it requires sacrificing task-level optimality. Across two distinct experimental paradigms, agents exhibiting patterned behavior were not only more predictable to their human partners, but were also perceived as more intelligent and cooperative. This work demonstrates that the key to effective human-autonomy teaming lies not in maximizing an agent's raw performance, but in optimizing its behavior for the cognitive frameworks of its human collaborators.

Two key findings provide insight into the cognitive mechanisms underlying this principle. First, we found that *patterned behavior enabled participants to form more accurate and stable mental models of the agent.* Their initial predictions required significantly less revision in the face of new information, providing quantitative evidence that human-aligned behavior fosters robust cognitive models. Second, *our results challenge the assumption that behavioral complexity serves as a proxy for perceived intelligence.* Participants judged the agent executing simple, patterned motion as more intelligent than the one executing mathematically optimal but opaque trajectories. This suggests that *for collaborative agents, the most salient feature of intelligence is not computational power or optimality, but the ability to make intentions transparent, thereby supporting the human's own cognitive processes.*

The embodied collaboration VR study further revealed that intuitive predictability is more critical for fluency than analytical transparency. Even when the environmental logic driving the optimal agent's behavior was visually apparent (i.e., avoiding patches of difficult terrain), participants were unable to form an effective predictive model. This demonstrates that simply making the reason for an agent's actions observable is insufficient. Fluent collaboration arises when the actions

themselves form a simple, perceptible pattern, engaging intuitive human cognitive processes rather than requiring conscious, analytical effort.

The principle of optimizing for human cognitive patterns has broad implications across robotics and artificial intelligence. In manufacturing, collaborative robots could use patterned motions to signal intent, allowing human workers to operate more safely and efficiently in shared spaces. In robot-assisted surgery, instruments could use conventional, patterned retraction movements to non-verbally communicate tool changes to the surgical team, improving workflow. For autonomous vehicles, encoding yielding intent into a perceptible braking pattern could resolve ambiguity for pedestrians, fostering trust and safety. More broadly, this work suggests a paradigm shift for AI design: from creating agents that are merely optimal to creating agents that are fluently compatible with human partners by explicitly anticipating that they are being observed.

Our findings should be interpreted in light of several limitations that suggest avenues for future research. The studies relied on university and online populations, and future work should validate these findings with more diverse, target user groups. Furthermore, the operationalization of patterns was tailored to navigation; future research must explore how to define and generate human-perceptible patterns in more complex, high-dimensional action spaces, such as those of manipulator arms. Finally, our experiments were conducted in simulation. Transferring this approach to physical systems will require addressing potentially confounding challenges borne of sensor noise, actuation error, and the need for robust online re-planning that preserves patterned structures in dynamic environments.

4.4 Materials and Methods

4.4.1 Human-Subjects Experimental Design

We conducted two human-subjects experiments to test how aligning autonomous behavior with human cognitive patterns affects team fluency. The first study investigated explicit prediction from a bird's-eye view (Online Study). The second measured implicit prediction and functional

utility in an embodied, co-located virtual reality task (VR Study). Both studies used a between-subjects design. Participants were randomly assigned to one of two conditions:

(1) **Optimal Behavior (control condition):** Participants interacted with a robot executing a reward-maximizing path generated by a reinforcement learning policy.

(2) **Patterned Behavior:** Participants interacted with a robot executing a path that was optimized for human-perceptible patterns, generated by post-processing the optimal path. The generation of these trajectory stimuli is detailed in Section 4.4.4.

4.4.2 Study 1: Online Experiment (Explicit Prediction)

4.4.2.1 Game Environment

The online study was conducted using the interface shown in Figure 4.2. Participants were shown part of the robot's trajectory and were asked to draw the remainder of the route they thought the robot would take to the goal. Participants were assigned a score for their guesses based on the distance from the actual trajectory. Participants completed ten rounds, making six predictions of the robot's trajectory each round. In each of the ten rounds, participants were first shown an empty environment, with marked start and goal locations, and instructed to draw the full path from start to goal as they thought the robot would behave. Then, the first $1/6^{th}$ of the robot's path was revealed, and participants were instructed to connect the visible portion of the trajectory to the goal area. This process was repeated, showing the participants an additional $1/6^{th}$ of the trajectory each time.

4.4.2.2 Protocol and Measurement

120 participants were recruited for the IRB-approved human subjects study on Prolific, an online platform for research studies. All users were fluent in English and based in the United States. Participants were able to join the study in only one of the two experimental conditions, which were labeled identically to prevent participants from knowing their assignment. We first obtained consent

from all participants. They answered pre-experiment survey questions taken from the Negative Attitudes towards Robots (NARS) scale to assess any pre-existing differences in attitudes towards robots. Between rounds, participants answered selected questions from the NASA Task Load Index (TLX) to measure their cognitive fatigue and frustration. The post-experiment survey consisted of questions from the Robotic Social Attributes Scale (ROSAS) to assess participant views about robot competence and likeability, survey questions about the fluency of the team, as well as custom questions about robot predictability and understandability. The duration of the experiment was approximately thirty minutes.

4.4.3 Study 2: VR Experiment (Implicit Prediction)

4.4.3.1 Game Environment

The VR environment was implemented using the Unity game engine and was deployed on the Meta Quest 3 VR headset. Participants collected rocks from the environment and brought them to the robot for analysis, while the robot navigated toward a marked goal across the map. This process can be seen in Figure 4.5. The robot moves continuously throughout the round so participants must implicitly predict where they think the robot will be in order to intercept it and deposit each rock. Participants engaged in five rounds of gameplay, each with a different, randomly selected environment. Each round of gameplay took approximately four minutes.

Environments were selected from the same bank used for the online experiment. A virtual environment was generated that indicated locations of obstacles and difficult terrain. States containing obstacles were mapped to impassable environmental features such as cliffs, lakes, and boulders. States of challenging terrain (areas of negative reward for the robot's policy) were mapped to steep slopes, tall grass, or gravel, where the robot would be forced to slow down.

4.4.3.2 Protocol and Measurement

Participants were recruited from the student community of our university for the VR study. Pre-experiment survey questions were taken from NARS and ROSAS, with the addition of demo-

graphics questions. Between rounds, participants answered questions from the NASA Task Load Index (TLX) to measure their cognitive fatigue and frustration. Participants were screened for cybersickness (motion sickness caused by virtual reality) before beginning each round to prevent them from participating while ill. The post-experiment survey consisted of questions from ROSAS, survey questions about the fluency of the team, and predictability questions available in the Appendix. Finally, participants engaged in a brief semi-structured interview with experimenters. The duration of the experiment was approximately sixty minutes.

4.4.4 Trajectory Stimuli Generation

A bank of trajectories were pre-generated for use within the two experiments. Each robot trajectory traversed a discretized 8-connected grid map produced from a bank of randomly generated environments, ranging in size from 5x5 to 20x20 grid squares.

4.4.4.1 Optimal (Reward-Maximizing) Trajectories

Reward-maximizing paths were computed using a universal policy across environments. This policy assigned high negative reward for states that were impassable to the robot and smaller negative rewards for areas that would slow the robot down due to the terrain type. Each action incurred a small negative reward to incentivize shorter paths.

4.4.4.2 Patterned (Human-Aligned) Trajectories

To generate the patterned stimuli, we applied a trajectory optimization post-processing layer, which we refer to as PRESTO (PREdictability-Satisfying Trajectory Optimization), to the reward-maximizing paths. This technique, which served as our instrument for generating predictable stimuli, generates paths by minimizing a weighted objective function that balances a predictability cost against a deviation cost using an optimal path as input:

$$\min_{\tau}(W \cdot C_{\text{predictability}}(\tau) + (1 - W) \cdot C_{\text{deviation}}(\tau, \tau^*))$$

where τ^* is the original reward-optimal reference trajectory, τ is the new candidate trajectory being optimized, and $W \in [0, 1]$ is a parameter that balances the trade-off between the two costs.

The Deviation Cost ($C_{\text{deviation}}$) penalizes candidate trajectories for divergence from the original reward-maximizing path, acting as a proxy for maintaining high task reward. It is defined as the summed pairwise dissimilarity between the states of the candidate and reference trajectories:

$$C_{deviation}(\tau, \tau^*) = \sum_{i=1}^{n} D(s_i, m_i)$$

where D is a problem-specific, non-negative State Dissimilarity Function. For our 2D navigation domain, we defined $D(s_i, m_i)$ as the Euclidean distance between the (x,y) locations of the state s_i in the candidate trajectory and the corresponding state m_i in the reference trajectory.

The Predictability Cost ($C_{\text{predictability}}$) quantifies how well the trajectory τ conforms to a clear, repeated pattern. Its calculation involves two steps:

(1) **Behavior Abstraction:** The state-space trajectory τ is mapped to a discrete symbolic behavior sequence $\gamma = [b_1, b_2, ..., b_{n-1}]$. This mapping first uses a human-perceptible feature extractor H to filter each state s to its observable features $\phi(s)$ — in our domain, this was the (x,y) location. A behavioral labeling function F_p then maps state transitions (s_i, a_i, s_{i+1}) to a discrete symbol b_i from a finite behavior alphabet B. For our 2D navigation domain, B was defined as a set of discretized directional changes. F_p computed the angle of change between successive waypoints (e.g., from the vector $w_{i-1} \rightarrow w_i$ to the vector $w_i \rightarrow w_{i+1}$) and "bucketed" the result into the corresponding discrete symbol in B.

(2) **Cost Calculation:** This cost is calculated from the symbolic sequence γ. We identify the longest repeated substring (LRS) α within γ, which represents the dominant repeating behavioral pattern. Letting k be the number of non-overlapping occurrences of α, the cost is formally defined as the number of behaviors in γ that do not belong to an instance of this primary pattern:

$$C_{\text{predictability}}(\tau) = |\gamma| - k \cdot |\alpha|$$

This cost function encourages the repetition of a fixed sequence of motions, rewarding trajectories composed of simple, repeating patterns.

The final optimization objective includes additional cost terms to enforce physical and environmental constraints. These terms penalized trajectories that: (i) passed through untraversable areas (e.g., obstacles, cliffs) ($C_{untraversable}$), (ii) placed waypoints outside the environment's defined bounds ($C_{boundary}$), or (iii) contained waypoints separated by more than the agent's maximum per-action movement distance ($C_{connectivity}$).For all patterned stimuli used in the experiments, the balancing weight W was tuned such that the final trajectory's reward loss (relative to τ^*) was bounded at no more than 20%. We selected environments using a pre-defined dissimilarity criterion to ensure the manipulation was salient and interpretable to participants; this may overestimate average-case benefits relative to truly random environments.

4.4.5 Statistical Analysis

Data were first assessed for normality using the Shapiro-Wilk test. For normally distributed dependent variables, between-group differences were analyzed using independent samples t-tests. For non-normally distributed data, such as Likert-scale survey responses anchored to the top or bottom of the scale, we used a Kruskal-Wallis test. Significant results were followed by a Dunn's post-hoc test with a Bonferroni correction for multiple comparisons to identify significant differences between the patterned and optimal conditions. All statistical tests used an alpha level of 0.05.

4.5 Conclusion

This work addresses a fundamental conflict between machine-centric optimization and human-centric cognition. Our findings across two distinct human-subjects studies, one measuring explicit prediction and the other measuring implicit, embodied collaboration, establish a core principle for human-autonomy teaming: aligning an agent's behavior with innate human cognitive faculties for pattern recognition yields substantial, measurable improvements in team fluency, efficiency, and

trust. We demonstrate that this alignment can be more critical for effective collaboration than pure algorithmic optimality.

By applying a trajectory optimization layer to inject human-perceptible patterns into task-optimal, reward-maximizing policies, we were able to systematically probe this principle. Our results showed that patterned behavior was not only more predictable, but it also reduced cognitive load, improved objective team performance, and was perceived as more intelligent than its task-optimal counterpart. This work provides quantitative evidence that patterned behavior fosters more stable human mental models, a cornerstone of effective team cognition. In our VR setting, terrain cues were insufficient to yield accurate prediction for the reward-optimal policy, whereas patterned motion supported prediction, revealing that intuitive predictability (seeing a simple pattern in actions) can be more effective for fluent collaboration than analytical transparency (seeing the environmental logic for an agent's optimal-but-complex actions).

This approach highlights the critical trade-off between machine performance and human predictability. While our method navigated this by bounding reward loss, it also reveals a limitation in using agent-centric "reward" as a universal currency. A human operator has no knowledge of the robot's reward function, making a bounded "reward loss" an abstract and often meaningless quantity in the context of a team task. Future work should focus on defining human-centric, rather than agent-centric, metrics to navigate this trade-off.

As autonomous systems become increasingly ubiquitous, the cognitive burden placed on human teammates by opaque, "optimal" behavior will become a critical barrier to adoption. This research provides evidence for a new design philosophy: shifting from optimizing for an agent's task to optimizing for a human's cognition. By fusing optimization methods with insights from cognitive science, we can create autonomous partners that are not just high-performing, but are also fundamentally aligned with the ways humans think, predict, and collaborate.

4.6 Extending Patterning to Multi-Agent Settings

In this work, we extend the idea of human-observable patterns into continuous spaces by first formalizing a method of identifying patterns within a trajectory, and validating it across multiple representations. Rather than limiting patterns to discrete task-level decisions, our approach deals with trajectories in the continuous space, making it applicable to more realistic human-robot settings. Importantly, the framework is designed to be generalizable across tasks and environments, allowing patterns to serve as a flexible structure for guiding robot behavior rather than a hand-crafted solution for a single scenario. We also introduce a principled method for balancing pattern adherence with reward maximization, enabling robots to remain as optimal as possible while improving predictability. Through validation across multiple human-robot collaborative scenarios, we show that this approach improves interaction quality and human perceptions while preserving performance. Additionally, the experiments indicate that perceived intelligence is not correlated with robot efficiency. We also see a significant increase in the stability of human mental models of robots, which have repercussions for future work.

However, all of these works involving patterning have thus far focused on settings with a single robot interacting with a person. This is a significant limitation, especially given what we know about human cognitive load in multi-agent settings. Cognitive demands can vary dramatically depending on coordination complexity, and existing evidence suggests that cognitive load increases substantially as the number of robots grows. To support more capable and realistic human-robot teams, we need to move beyond single-robot interactions and examine how patterns function in multi-robot contexts. This next step is critical for understanding how to design coordinated robot teams that remain predictable, manageable, and effective for human partners in more complex environments.

Chapter 5

Pedestrian-Inspired Patterning as Structural Compression in Human–Robot Teams

The previous work shows that it is possible via the use of patterns to strike a meaningful balance between predictability and optimality in robot behavior. By explicitly incorporating patterns into planning, robots can behave in ways that are easier for people to understand without completely sacrificing efficiency or task performance. This balance is especially important in collaborative settings, where humans must quickly interpret robot actions and adapt their own behavior accordingly. However, even when individual robot behavior is more predictable, the cognitive demands placed on the human teammate can still be substantial. This challenge becomes even more pronounced as the complexity of the interaction increases, especially when additional agents are in the shared environment.

Multi-agent settings introduce a significant source of cognitive load, as people must model, monitor, and coordinate with multiple robots at once. Adding more agents can quickly make the system harder to understand and manage, even if each robot is individually predictable. Pedestrian modeling has already studied the interactions between humans navigating in crowds, providing useful insights into how people naturally coordinate in these environments. This raises an important opportunity: rather than treating multi-agent navigation as a purely optimization problem, we can ask whether these existing human-derived patterns can be leveraged to guide groups of multiple robots. Doing so could help make larger robot teams more predictable and easier for people to work with in complex collaborative tasks.

5.1 Introduction

Multi-agent human–robot systems are increasingly being deployed in real-world environments such as public spaces, warehouses, transportation hubs, and collaborative work settings. In these domains, humans are often required to observe, interpret, and anticipate the behavior of many autonomous agents simultaneously. Scaling human understanding to dense interactive multi-agent systems remains a fundamental challenge. As the number of agents increases, the cognitive demands placed on human observers grow rapidly, limiting their ability to form accurate predictions and make timely decisions. A central difficulty in these settings is not only the complexity of individual agent behavior, but *the combinatorial growth of relational structure*. In multi-agent systems, behavior is defined not just by isolated trajectories, but by interactions, coordination, and group dynamics. This creates a representational burden for human observers, who must track not only what each agent is doing, but how agents collectively organize and evolve over time. As system complexity increases, humans are forced to rely on simplified internal representations and heuristics [13].

Prior work in human–robot interaction has primarily focused on improving interpretability through local mechanisms, such as making individual agents more predictable, increasing transparency of intent, or increasing communication [31, 71, 72, 30, 93, 49, 100, 43, 105, 21, 97]single-agent settings, they often treat agents as independent units and do not directly address how humans construct higher-order representations of multi-agent systems. As a result, they overlook a critical aspect of human understanding: the ability to simplify complex multi-agent behavior into structured, meaningful patterns that operate at a level of abstraction above individual agents[13].

In this work, we propose a different perspective grounded in what we define as structural compression: the reduction of a multi-agent system's effective representational complexity through the emergence of higher-order structures, such as dynamically formed groups. We argue that human performance in multi-agent prediction tasks depends not only on reducing uncertainty at the level of individual agents, but also on enabling observers to form compact, higher-order mental models of system behavior. Specifically, we introduce pedestrian group models as a mechanism for

inducing structured organization in multi-agent systems, allowing behavior to be mentally compressed into group-level dynamics. Rather than treating behavioral complexity as something to be directly minimized, we explore how introducing interpretable structure through pedestrian-inspired coordinated grouping and familiar behavioral patterns can actively support human cognition by objectively reducing effective representational complexity.

We make the following contributions. First, we introduce structural compression as a conceptual framework for understanding human performance in multi-agent interaction, emphasizing the role of representational efficiency over purely perceptual or workload-based explanations via the use of computed metrics. Second, we propose and evaluate a pedestrian-inspired decentralized grouping algorithm that dynamically structures agent behavior to support group-level interpretation. Third, we demonstrate empirically that coordinated grouping selectively improves human predictive accuracy in high-complexity conditions, where the benefits of compressed representations become most pronounced. Finally, we show that these performance improvements are not consistently reflected in subjective measures such as perceived workload or understandability, suggesting a dissociation between experiential and representational effects. Our results suggest that effective support for humans in large-scale multi-agent systems may require moving beyond individual-agent design and toward the intentional structuring of collective behavior. By shaping how complexity is organized rather than simply reducing it, system designers may better align external system dynamics with the internal representational strategies humans naturally use to manage complexity to enable more effective collaboration.

5.2 Background and Related Work

Our work builds on collective emergent behavior, social force models, and existing pedestrian group models in order to produce agentic behavior that mimics pedestrian grouping behavior.

5.2.1 Collective Emergent Behavior

Emergent behavior refers to complex, coordinated patterns that arise from the local interactions of individual agents following relatively simple rules[25]. In crowd dynamics, these behaviors, such as spontaneous lane formation, collective turning, and synchronized flow through bottlenecks do not require explicit global control or communication. Instead, they emerge organically as each individual continuously adapts to the movements and intentions of others. This bottom-up organization is a hallmark of many natural and social systems, highlighting how structure and predictability can result from decentralized decision-making[25].

Emergent behavior is closely tied to human perceptual and cognitive abilities to detect and respond to patterns in their surroundings [69, 25, 48]. As pedestrians navigate shared environments, they continuously interpret subtle cues such as speed, gaze direction, and interpersonal spacing, to infer others' intentions and adjust their own motion accordingly [48]. This innate patterning ability allows individuals to synchronize movements without explicit communication, reinforcing collective structures like walking lanes and flow partitions. In turn, these emergent patterns reduce cognitive load and enhance overall efficiency by creating predictable pathways through dynamic crowds [48, 13, 8, 87]. The interplay between human pattern recognition and emergent behavior underscores the importance of designing autonomous systems that not only avoid collisions but also engage with the implicit social rhythms that guide human group motion, which we explore in this work.

5.2.2 Social Force Models

Social force models are a widely adopted theoretical framework for representing pedestrian and crowd dynamics through mathematically defined "forces" governing agent motion[48]. In this formulation, each pedestrian is treated as a self-driven entity with a preferred velocity toward a goal, while additional repulsive and sometimes attractive forces modulate interactions with other agents and the environment. These forces are not literal physical interactions but abstractions that

encode behavioral tendencies such as maintaining personal space, avoiding collisions, and adhering to spatial conventions. By embedding these behavioral principles into continuous optimization of trajectory and velocity, social force models enable systematic simulation and analysis of navigation behaviors at both local and global scales.

Crucially, social force models facilitate the emergence of complex collective phenomena from simple interaction rules. Empirical studies have demonstrated their ability to replicate observed pedestrian behaviors, including lane formation in bidirectional flows, group cohesion among socially connected individuals, and non-linear congestion patterns in constrained environments[48]. Extensions to the foundational model further incorporate elements such as predictive collision avoidance, heterogeneous population characteristics, and context-specific behaviors relevant in emergency or high-stress scenarios [112, 29, 70, 26].

5.2.3 Extended Social Force Models - Groups

Social force models that explicitly incorporate group dynamics extend the foundational framework by accounting for the social bonds and shared goals that influence pedestrian behavior[88]. Rather than treating each individual as an isolated agent, these models introduce group cohesion and alignment forces that preserve proximity and coordinated movement among members of the same group [88]. Such forces capture underlying social motivations and ensure that group members adjust their trajectories in ways that maintain collective identity while still respecting collision avoidance constraints. This enriched structure allows simulations to reflect more realistic, heterogeneous populations where interactions vary significantly between intra-group and inter-group encounters[112, 124, 121].

These group-aware extensions are particularly important for understanding how collective behavior shapes crowd-level phenomena. For example, cohesive groups tend to move more slowly and occupy more space, creating flow disturbances and bottlenecks in tightly constrained environments [112, 124, 121]. Group models can also capture phenomena such as group splitting and merging, leadership dynamics, and the influence of strong social ties during evacuation scenarios

[112]. In applied contexts such as autonomous navigation and human-robot interaction, leveraging group-sensitive social force models enables robots to interpret and respond to group formations more appropriately, whether by integrating into existing clusters or avoiding disruption of social units [124]. In this way, group-oriented social force frameworks provide a more nuanced and human-centered foundation for analyzing and shaping motion in shared environments. This group-oriented social forces model serves as a foundation to our method.

5.2.4 Multi-Agent HRI

Multi-agent human–robot interaction (HRI) has received increasing attention as robotic systems move from isolated single-robot settings to dense, shared environments involving many autonomous agents[32, 74]. Compared to dyadic interaction, multi-agent settings introduce substantially greater perceptual and cognitive demands, as humans are required to monitor, interpret, and often predict the behavior of multiple interacting entities simultaneously. A central challenge identified in this literature is the rapid escalation of cognitive load with increasing numbers of agents, driven not only by the need to track more objects, but also by the added complexity of interactions among them[59, 32, 74]. As a result, human performance in tasks such as monitoring, prediction, and decision-making tends to degrade as system scale and interaction density increase, highlighting fundamental limits in attentional capacity and working memory [13].

To address these challenges, prior work has focused on reducing cognitive load through improvements in individual agent design and system transparency[32, 10]. In particular, research in HRI and related areas such as swarm robotics has explored methods for making agent behavior more predictable, legible, and easier to track[59, 32]. Techniques include motion planning that prioritizes human interpretability, explicit communication of intent, and visualization methods that externalize agent state or predicted trajectories[35, 23]. These approaches are often motivated by cognitive load reduction, with the assumption that making individual components easier to understand will reduce overall mental effort in multi-agent tracking[59]. Similarly, work on human supervision of robot teams has investigated abstraction mechanisms and interface designs intended

to offload tracking demands and support more efficient allocation of attention[32, 74].

However, despite these advances, existing approaches focus on reducing cognitive load at the level of individual agents rather than addressing how humans represent multi-agent systems as structured wholes[35]. While some work acknowledges that humans may rely on heuristics such as grouping or summarization under high load, these mechanisms are typically treated as emergent phenomena rather than explicit design targets. As a result, relatively little attention has been given to how designers can shape system-level structure to align with human representational strategies. This leaves an open gap in understanding how multi-agent systems might be designed not only to reduce local uncertainty or attentional demand, but also to support more efficient global representations of collective behavior, particularly under conditions where full individual tracking is infeasible. This work addresses this gap by introducing pedestrian-inspired dynamic grouping to imposed structure over the system as a whole, allowing for more efficient global representations.

5.2.5 Predictability in HRI

Across HRI, and especially in multi-agent settings, the ability to accurately predict an agent is deeply important as predictability is strongly correlated with trust, team fluency, and positive perceptions of robot teammates [36, 71, 31, 72, 27]. In this work, we define predictability as "the quality of matching expectations" [31]. The expectations that a human has of a robot are derived from their mental model of the robot. Mental models are structures that humans build in their minds to help navigate environments, make decisions, and reason about collaborators [108]. Humans are exceedingly skillful at constructing mental models about other human collaborators [122], and the more accurate a human's mental model of another is the better they are able to collaborate with each other [79]. This concept translates into human-robot interaction, as humans also construct mental models of the robots that they work with [108, 71], and the more accurate the human's mental model of the robot is, the more effective the human-robot collaboration will be [90, 56, 125, 79]. As these mental models encode human expectations, creating robot behaviors that are congruous with these human mental models (e.g., creating autonomous systems that act predictably) is criti-

cal for effective human-robot collaboration [65, 31, 27, 71, 36]. This work seeks to directly encode human expectations of group motions into agentic behavior in order to strengthen human mental models of robots, as well as improve agent predictability.

5.3 Methodology

In our approach, we utilize the extended social force model for groups introduced by Moussaid et. al. to constrain agentic behavior to mimic pedestrian groups. We adapt this model to be dynamic, such that agents can form or dissolve groups at a given cadence. Additionally, we modify this model to be decentralized, such group mechanics can be calculated by an individual agent without total knowledge of the entire group. Further, our decentralized method allows for robots to align themselves with mixed human-robot groups, who are not able to participate in the algorithm, but whose movement can be observed by an agent. A walkthrough of the methodology at a high level can be seen in Figure 5.1.

5.3.1 Social Force Model

The classic Social Force Model (SFM), introduced by Helbing and Molnár [48], describes each agent i as a self-driven particle whose motion evolves according to a set of continuous forces. The core equation governs the change in velocity for each agent over time, $\frac{d\vec{v}_i}{dt}$, as a relaxation toward a desired velocity while satisfying social constraints is:

$$\frac{d\vec{v}_i}{dt} = \vec{f}_i^{0} + \vec{f}_i^{\mathbf{wall}} + \sum_{j \neq i} \vec{f}_{ij}^{\text{soc}} \tag{5.1}$$

Here, $\vec{f}_i^{0}$ is the driving force of the agent i toward the agent's goal, $\vec{f}_i^{\mathbf{wall}}$ is the repulsive force from obstacles, and $\sum_{j \neq i} \vec{f}_{ij}^{\text{soc}}$ is the sum of all repulsive forces between all other agents. The driving force enforces convergence to a preferred speed and direction:

$$\vec{f_i^{0}} = \frac{d\vec{f_i}}{dt} = \frac{v_i^0 \vec{e_i^0} - \vec{v_i}(t)}{\tau} \tag{5.2}$$

where $\vec{e_0^i}$ is the desired velocity vector and τ is a relaxation parameter representing how quickly the agent corrects deviations from its intended motion. v_i^0 is agent i's desired speed, which is a value

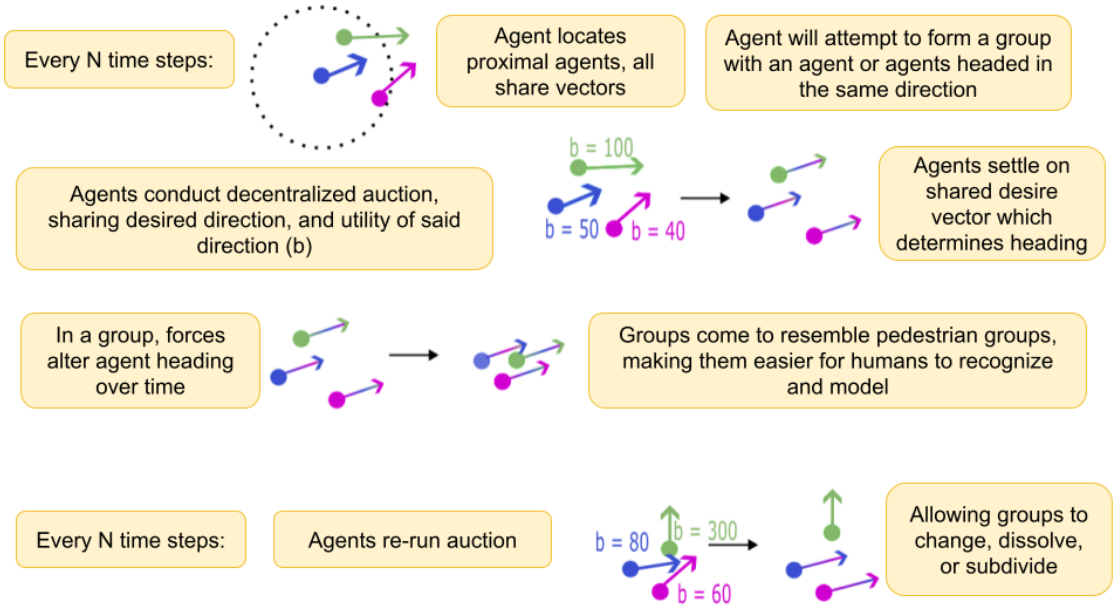

Figure 5.1: **Conceptual illustration of the intervention process.** Agents initially move independently toward individual goals, requiring observers to track multiple distinct trajectories. The intervention identifies subsets of agents with sufficiently aligned motion and groups them into higher-order units based on shared directional structure. This intervention also changes how close agents get to another, and what shape the group takes. By representing coordinated agents as collective patterns rather than separate entities, the system reduces the effective number of agents that must be considered, enabling a more compressed and tractable representation of complex multi-agent behavior.

empirically derived in prior work, and $\vec{v_i}(t)$ is the agent's velocity.

Repulsive forces from obstacles are modeled by:

$$f_i^{\vec{\textbf{wall}}}(d_w) = ae^{-\frac{d_w}{b}} \tag{5.3}$$

Where d_w is the distance to the obstacle, and a and b are empirically determined constants for the strength and range of the repulsion.

Interpersonal forces model collision avoidance and personal space maintenance using an exponentially decaying potential. The repulsive component between pedestrians i and j is defined as:

$$\vec{f_{ij}} = [A_i^{\frac{r_{ij}-d_{ij}}{B_i}} + kg(r_{ij} - d_{ij})]\vec{n_{ij}} + \kappa g(r_{ij} - d_{ij})\Delta v_{ji}^t \vec{t_{ij}} \tag{5.4}$$

where A_i controls interaction strength, B_i controls interaction range, $r_{ij} = r_i + r_j$ is the sum of body radii, d_{ij} is the current inter-center distance, and $\vec{n}_{ij}$ is the normalized vector pointing from pedestrian j to pedestrian i. k is the body stiffness constant, with higher stiffness leading to stronger pushing forces. Tangential components may be added to model friction-like effects during near-contact, enhancing realism in congestion (κ is the tangential friction coefficient that controls sliding resistance).

5.3.2 Extended Social Force Model for Groups

This work utilizes the extended force model for groups, as define by Moussaïd et. al[88]. This extension accounts for static groups of pedestrians, modeled after real-world data. This model extends the prior social force model as such:

$$\frac{d\vec{v_i}}{dt} = \vec{f_i^0} + \vec{f_i^{\textbf{wall}}} + \sum_{j \neq i} \vec{f_{ij}^{\textbf{soc}}} + \vec{f_i^{\textbf{group}}} \tag{5.5}$$

$\vec{f_i^{\textbf{group}}}$ is composed of three aspects of group forces:

$$\vec{f_i^{\textbf{group}}} = \vec{f_i^{\textbf{vis}}} + \vec{f_i^{\textbf{att}}} + \vec{f_i^{\textbf{rep}}} \tag{5.6}$$

These components are forces to keep the group within visual field ($\vec{f}_i^{\mathbf{vis}}$), an attraction toward the group's center of mass ($\vec{f}_i^{\mathbf{att}}$), and a repulsive force such that group members do not overlap ($\vec{f}_i^{\mathbf{rep}}$). $\vec{f}_i^{\mathbf{vis}}$ will be omitted in this work, as this is purely to accommodate conversation, which is not a behavior the robots are engaging in within this context. Group attraction force is defined as follows:

$$\vec{f}_i^{\mathbf{att}} = q_A \beta_2 \vec{U}_i \tag{5.7}$$

where $q_A = 1$ if the distance from the agent to the center of mass is greater than a given threshold, and $q_A = 0$ otherwise. β_2 is the strength of the attraction effects, and $\vec{U}_i$ is the unit vector pointing from pedestrian i to the center of mass of the group.

Group repulsive force is defined as:

$$\vec{f}_i^{\mathbf{rep}} = \sum_k q_R \beta_3 \vec{W}_{ik} \tag{5.8}$$

where $\vec{W}_{ik}$ is the unit vector pointing from pedestrian i to the group member k and β_3 is the repulsion strength. $\vec{W}$ is a group-level directional influence vector, not merely the vector between two agents. $q_R = 1$ if pedestrians i and k overlap each other (when the distance d_{ik} is smaller than a threshold value d_o, that is one body diameter plus some safety distance), otherwise $q_R = 0$. The desire vector calculated via our method can be converted into a heading to be used by an outside navigation or planning algorithm.

5.3.3 Dynamic Joining and Leaving Groups

The prior work by Moussaid et al [88] is predicated upon static groups. Agents start in a group and remain in the same group over time. To utilize this model as-is, agents would have to share information about themselves such as goal locations. However, this would require divulging information to unknown agents in more realistic environments, which is not desirable, so we propose a decentralized method for dynamic groupings. Additionally, our decentralized method allows for robots to align themselves with mixed human-robot groups, who are not able to participate in the algorithm, but whose movement can be used by the robots to "tag along" with humans.

In order to maximize predictability, it would be more advantageous for agents to dynamically group and ungroup based on shared direction in the moment. When humans are in a crowd, they may follow in the wake of a larger group to make navigation easier. Amongst agents, those agents traveling a great distance may find their desired directions to be somewhat aligned in the short term, and group together until their desires diverge when closer in proximity to their goals. This allows for a human observer to see that all agents in a clump are moving in a shared general direction, making them easier to predict.

5.3.3.1 Group Joining Method

Individual agents will plan and move according to any algorithm selected by a designer, and grouping adds on to these methods by giving the outside methodology a heading that accounts for the agent's goal as well as helping it to engage in grouping behavior. Every t steps, an individual agent a_i will identify a set of neighbors, N, that surround the agent within a given radius r. All agents will broadcast their velocity vector $\vec{v}_i$. These values could be derived from visual data as well (eg, in mixed human-robot groups). If the vectors are sufficiently aligned, the definition of which is set by the designer, the agent will become part of a group. If the neighbor agent is in a group already, the agent will join its group, and if not, the aligned agents will form a new group.

Groups will then engage in a decentralized auction to determine the desire vector, $\vec{e_i}$ for the group. Each agent's initial candidate vector $\vec{c_i}$ will be obtained from whatever method of planning the agent is using. This value will be converted from the planning algorithm's desired heading. The agent's bid (b_i) with their vector, will be equal to:

$$\frac{1}{d_{i\rightarrow\mathbf{goal}}}$$

where $d_{i\rightarrow\mathbf{goal}}$ is the distance from agent i to its goal. Thus, an agent is less willing to deviate from the optimal path the closer it is to the goal.

Each agent will update its candidate vector using a weighted average of candidates and bids within its neighborhood. Candidate vectors will be updated until $|c_i - c_i'| < \epsilon$, where c_i is the

original candidate vector and c_i' is the updated candidate vector. The final value will be used as a desire vector in the social forces model, the group forces calculations, and the resulting vector will be converted to a heading to be used by the agent's planning algorithm.

5.3.3.2 Group Leaving Method

There are many methods of determining a group is losing cohesion, including diverging desire vectors, increasing social forces, and agent dispersion. However, these values are not necessarily stable with optimized agents. Every t timesteps, groups will be re-assessed via the same decentralized auction used to create groups. The same decentralized auction allows for the creation of diverging subgroups, or for individual agents to leave all groups.

5.3.4 Desire Vector and Theta Conversion

While a traditional desire vector as formulated by Helbing et. al. can be calculated for any agent given the agents surrounding it and its goal, an appropriate heading can also be obtained from an outside planning method and converted to a pseudo-desire vector, which can then be used to engage in grouping behavior. The resulting vector from the decentralized auction can also be converted back to a θ value to be used by an outside planner or navigational algorithm. This work assumes directional changes as immediate, with updates to agents' headings not taking any timesteps.

5.3.5 Effective Agent Count

To quantify the structural changes introduced by the grouping intervention, we define an effective agent count N_{eff} that captures the number of distinct groups present in the system at a given prediction timestep. N_{eff} reflects the number of agents that would need to be tracked if groups are represented as a single agent.

In the coordinated grouping condition, agents self-identify as members of a group through the decentralized auction process. At regular update intervals, groups may form, dissolve, or subdivide

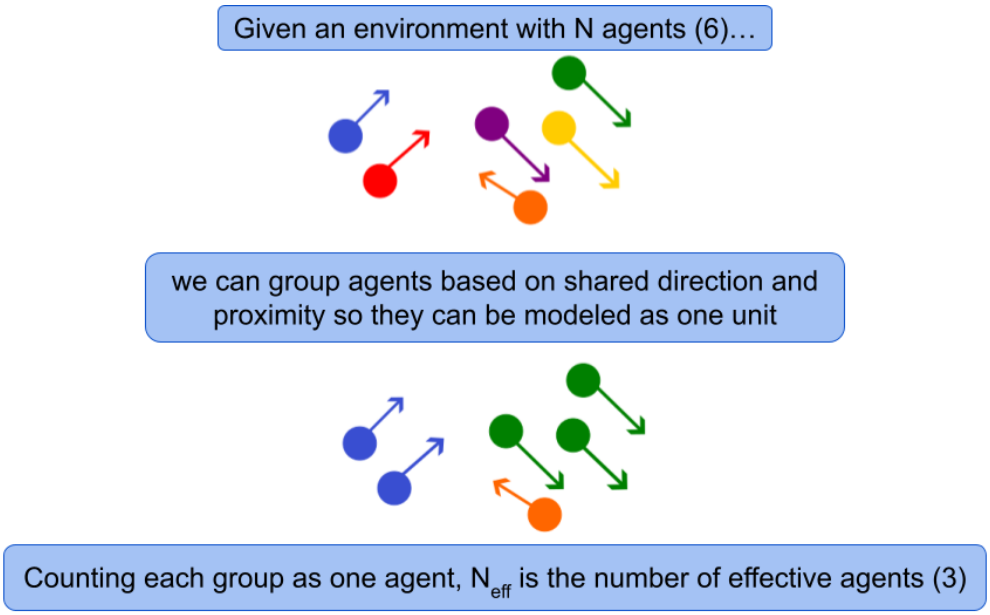

Figure 5.2: **Conceptual illustration of effective agent count (N_{eff}) as a measure of structural compression.** Agents moving in sufficiently aligned directions form decentralized groups, allowing multiple physical agents to be represented as a single functional unit. N_{eff} is defined as the number of such active groups at a given timestep. A lower N_{eff} indicates greater structural compression, reflecting a reduced number of effective units required to represent system behavior. This metric provides a normalized basis for comparing complexity across conditions and prediction intervals.

depending on local motion alignment and auction outcomes. For each timestep corresponding to a participant prediction event in the experiment, N_{eff} was recorded as the number of active groups plus the number of ungrouped agents. This value therefore represents the instantaneous structural complexity of the system under the intervention. An example of the differences between N and N_{eff} can be seen in Figure 5.2.

For the independent-agent baseline, no explicit grouping mechanism exists. To enable comparison across conditions, N_{eff} was estimated by identifying clusters of agents whose trajectories exhibited sufficient directional similarity at the same sampled timesteps. This produced a baseline measure of grouping that is directly comparable to the intervention-derived group count, while preserving the absence of any explicit coordination mechanism. Even in the baseline case, N_{eff} is not always equal to N, especially at the start or end of motion, or when agents must navigate through constrained spaces, as the threshold for being considered sufficiently grouped is met, though the shape of the groups does not match those of pedestrian groups.

5.3.6 Compression Ratio and Compression Gain

To normalize across rounds with different total numbers of agents (N), we further compute a compression ratio:

$$\textit{Compression Ratio} = \frac{N_{eff}}{N} \tag{5.9}$$

and its complementary compression gain:

$$\textit{Compression Gain} = 1 - \frac{N_{eff}}{N} \tag{5.10}$$

Under this definition, a higher compression gain indicates that the observable system can be represented using fewer effective units relative to the total number of agents. These metrics provide a common quantitative framework for evaluating how strongly the intervention reduces the dimensionality of the prediction problem at each participant decision point.

5.4 Experimental Validation

5.4.1 Experimental Environment

In this study, we assess participants' ability to predict multiple agents' behavior by asking them to draw the paths they would take to their goals. The study was conducted using the interface shown in Figure 5.3, as it is necessary to prove the baseline cognitive phenomenon via direct prediction prior to conducting further experiments. Each agent is assigned a color; the agent is depicted as a dot, and its goal as an 'X'. Obstacles are depicted as black circles and ovals. Participants were shown part of the robots' trajectories and were asked to draw the remainder of the route they thought the robots would take to their goals, labeled with Xs. Participants completed ten rounds with a different agent and environmental configuration for each, making six predictions of the robots' trajectories each round. In each of the ten rounds, participants were first shown an empty environment, with marked start and goal locations for each agent, and instructed to draw the full paths from start to goal as they thought the robots would behave. Then, the first $1/6^{th}$ of the robot's path by time was revealed, and participants were instructed to connect the agent to the goal area again, using the visible previous part of the trajectory to assist in their prediction. This process was repeated, showing the participants an additional $1/6^{th}$ of the trajectory each time. The first round began with only one agent being predicted, with the number of agents increasing across rounds, with a final count of nine agents in the tenth round. Agent counts for each round are as follows: 1, 2, 3, 3, 4, 5, 6, 6, 7, and 9. Agent counts were fixed across all participants, but environments, obstacles, and agent paths were selected randomly from a precomputed bank of possible environments. The first round with only a singular agent was used as an introductory round to allow participants to familiarize themselves with the activity.

5.4.2 Experimental Design

Study participants (n = 21, total of 1,200 observations collected) were randomly assigned into one of two groups, which determined the methodology used for agent motion.

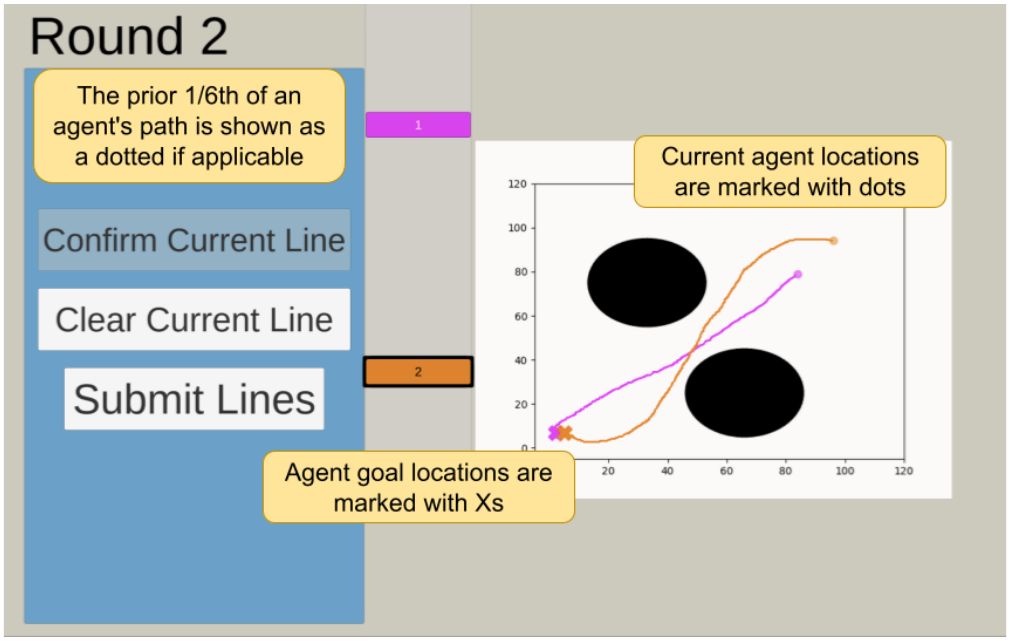

Figure 5.3: **A screenshot from the user interface of the study.** Participants must predict each robot's path by drawing on the environment. Each round consists of a sequence of trials that progressively reveal more of the robots' actual paths from start to goal (color-coded) in 16.67% increments, beginning with none of the path revealed. Participants draw a path prediction predicting the completion of the path.

- Unstructured Agents (Baseline) - agent motion will be controlled by an RRT* planner.
- Coordinated Groups (Experimental) - agent paths will be the baseline paths with dynamic grouping altering agent behavior when indicated by the algorithm.

5.4.3 Study Protocol

21 participants were recruited for the IRB-approved human subjects study on Prolific, an online platform for research studies. All participants were fluent in English and based in the United States. Participants were able to join the study in only one of the two experimental conditions, which were assigned randomly in Qualtrics. The duration of the experiment was approximately forty-five minutes.

5.4.4 Measurement

Prior to the experiment, participants were asked about their experience and opinions of AI and robots, using Likert-scale questions from prior work [91], to account for any baseline differences between groups. After each round, participants answered the full NASA TLX battery[44], and two questions about their ability to predict the agents. The post-activity survey was comprised of questions about the predictability and understandability of the agents, as well as participants' perceptions of the agents, taken from the RoSAS scale[16] and prior teaming work[50].

Quantitative metrics of performance were also collected, including direct prediction accuracy (RMSE), time taken, and logging of all participant actions taken within the game environment. RMSE was calculated on a pointwise basis between the ground-truth agent location and the participant-predicted location. When exact matching x-values were unavailable, linear interpolation between adjacent participant-predicted points was used to align samples. Because participant prediction drawings were continuous and densely sampled, the resulting interpolation introduced minimal distortion. Furthermore, by capturing continuous trajectory data rather than discrete prediction outcomes, this measurement approach yields a high-resolution, dense volume of localized prediction errors per trial, maximizing the statistical power derived from the participant pool.

5.4.5 Hypotheses

- H_1: Coordinated grouping will allow humans to more accurately predict more agents than a baseline planner.
- H_2: Coordinated grouping will improve the perceived predictability of agents based on participant self-reported perceptions between rounds and post-activity in provided surveys.
- H_3: Coordinated grouping will result in lower cognitive load for humans managing multiple agents as measured by self-reported NASA-TLX scores participants provide after each round.

5.5 Results

Of the 21 individuals who participated in our IRB-approved study, the data of one participant was excluded due to noncompliance with instructions. We did not observe any multimodalities within the data. No significant differences were observed between groups in the pre-activity survey. Post-hoc comparisons of the data were conducted with the Mann-Whitney U test.

5.5.1 H_1: Objective Performance

H_1 proposed that the coordinated grouping would improve participants' quantitative performance, as reflected in objective accuracy measures. Performance was evaluated using RMSE across rounds and task conditions.

The results provide partial support for this hypothesis. Participants in the pedestrian-inspired groups condition demonstrated improved prediction accuracy compared to the baseline condition; however, this effect was not consistent across all rounds ($p = \mathbf{.049}, p = .83, \mathbf{p < .001}, p = .31, p = .36, p = .33, \mathbf{p = .02}, .6, \mathbf{p < .001}, \mathbf{p < .001}$). To account for repeated statistical testing across rounds, Bonferroni correction was applied to all reported p-values. Instead, the advantage emerged most clearly in rounds involving higher numbers of agents, where the prediction task was more

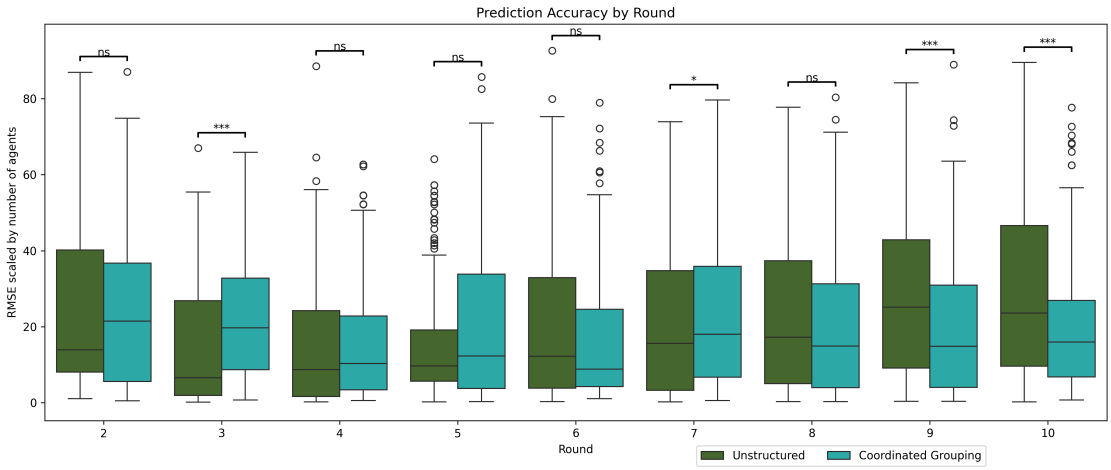

Figure 5.4: **Coordinated grouping provides inconsistent reductions in root mean square error (RMSE).** Differences between groups were inconsistent during earlier rounds, which involved fewer agents and lower task complexity, with significance appearing only sporadically. In contrast, later rounds (characterized by higher agent counts) showed more consistent and sustained significant differences, indicating that group effects became more pronounced as scenario complexity increased. This pattern suggests that the intervention's benefit emerged selectively under higher-complexity conditions rather than uniformly across all rounds.

complex (Round 7: $p = .02$, Round 9: $p < .001$, Round 10: $p < .001$). Additionally, no significance was found in the time participants took to complete the task.

As seen in Figure 5.4, in rounds with fewer agents, the effect was less stable and in some cases absent entirely. This suggests that the benefits of the experimental approach were not uniform across task demands. This indicates that the intervention was most effective under conditions of increased cognitive complexity, where participants were required to track and anticipate the behavior of a larger set of agents.

Overall, these findings suggest that the use of coordinated, pedestrian-inspired groups can enhance human predictive performance, but its effectiveness depends on task context. The hypothesis is therefore only partially supported, with benefits concentrated in scenarios involving greater agent counts rather than across the full range of scenarios.

Taken together, these findings suggest a dissociation between perceived and actual system performance: while participants did not rate the system as more predictable or understandable than the baseline group, their behavioral outcomes nevertheless improved under specific task conditions.

5.5.2 H_2: Perceived Predictability and Understandability

H_1 proposed that improvements in system predictability and understandability would be reflected in participants' subjective assessments across rounds and in post-task evaluations. To examine this, two questions posed after each round as well as ten questions from the post-activity survey were analyzed. Questions related to predictability and understandability in the post-activity survey were summed and scaled by the number of questions to create two superscore measures capturing participants' perceived predictability and understandability($\alpha = .86$).

Overall, the hypothesis was not supported. Subjective ratings did not show meaningful differences between groups; participants did not report higher confidence in their predictions between any of the rounds (all $p > .06$), nor did they rate the agents and being more predictable ($p = .499$) or understandable ($p = .53$) after the activity, as seen in Figure 5.5.

However, this pattern contrasts with the quantitative performance metrics, as discussed pre-

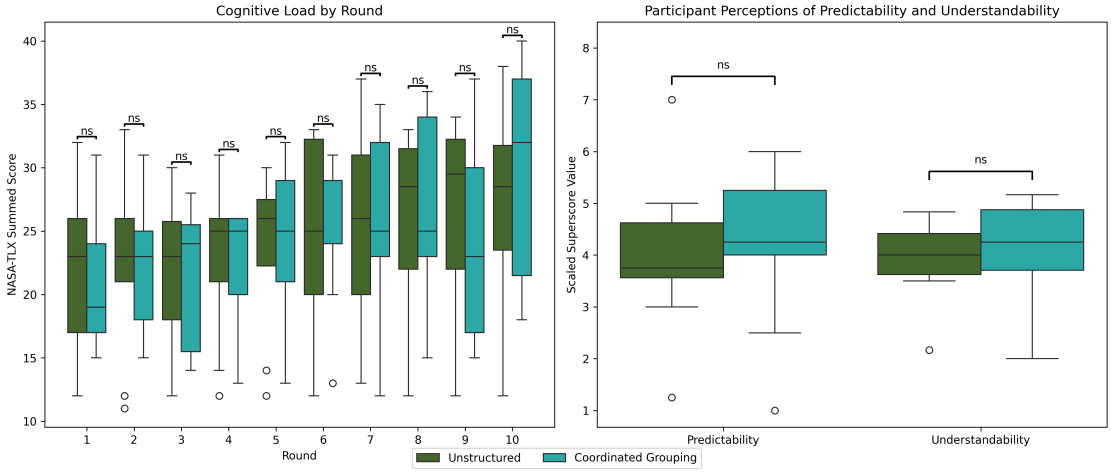

Figure 5.5: **Absence of subjective effects between conditions**. Left: Cognitive load ratings across rounds for the independent-agent (baseline) and coordinated grouping (experimental) conditions. No statistically significant differences were observed at any round, indicating that the coordinated grouping did not measurably alter perceived task effort at any point. Right: Post-task self-reported measures of predictability and understandability. No significant differences were found between conditions, suggesting that despite differences in objective performance and system structure, participants did not self-report changes in subjective perceptions of predictability or understandability of the multi-agent system.

viously. The objective error-based measure used (RMSE) indicate that participants' performance improved in certain rounds despite the lack of subjective improvement, particularly in conditions with higher agent counts, as seen in Figure5.4. This contrast between objective and subjective metrics is often uncovered in HRI work, so this finding is not unexpected.

5.5.3 H_3: Cognitive Load

H_2 proposed that participants in the Coordinated Grouping condition would report lower cognitive load than those in the baseline condition, as measured by the NASA-TLX. To evaluate this, we compared NASA-TLX scores between conditions and across rounds, the results of which can be seen in Figure 5.5.

Contrary to this hypothesis, the results showed no significant differences in cognitive load between the experimental and baseline groups. Across all rounds, NASA-TLX scores remained statistically indistinguishable, indicating that the experimental intervention did not reduce perceived workload relative to the baseline condition (all $p > .46$).

This null effect was consistent throughout the study: neither repeated exposure nor condition differences led to measurable changes in subjective cognitive load, as seen in Figure5.5. Participants reported similar levels of effort regardless of group assignment, suggesting that the pedestrian-inspired grouping did not meaningfully impact perceived workload even as other performance measures varied.

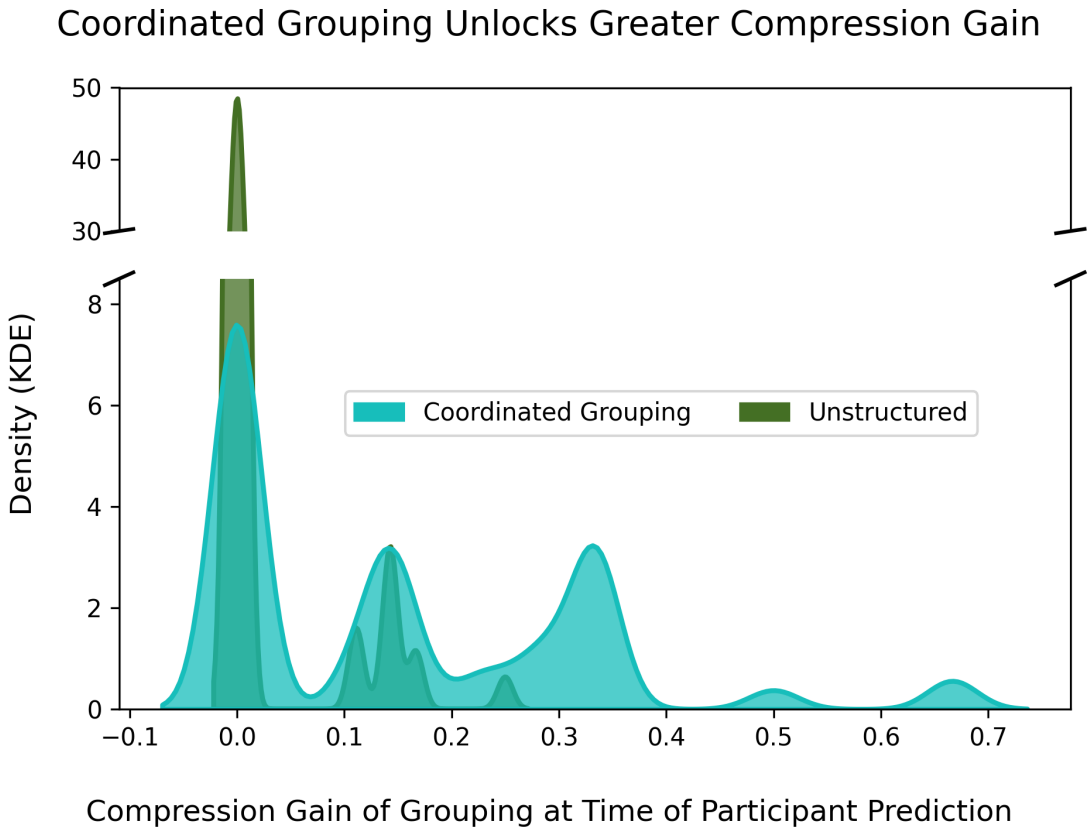

Figure 5.6: **Pedestrian-inspired grouping unlocks greater compression gain.** Compression gain quantifies the proportional reduction in effective agent count relative to the total number of agents. The independent-agent condition is concentrated near low compression values, indicating that most observations retain a high effective complexity. In contrast, the coordinated grouping condition exhibits a broader distribution and extends into substantially higher compression regimes. This shift demonstrates that the intervention does not merely reduce average complexity, but enables access to representational states characterized by greater structural compression than those possible in the baseline condition.

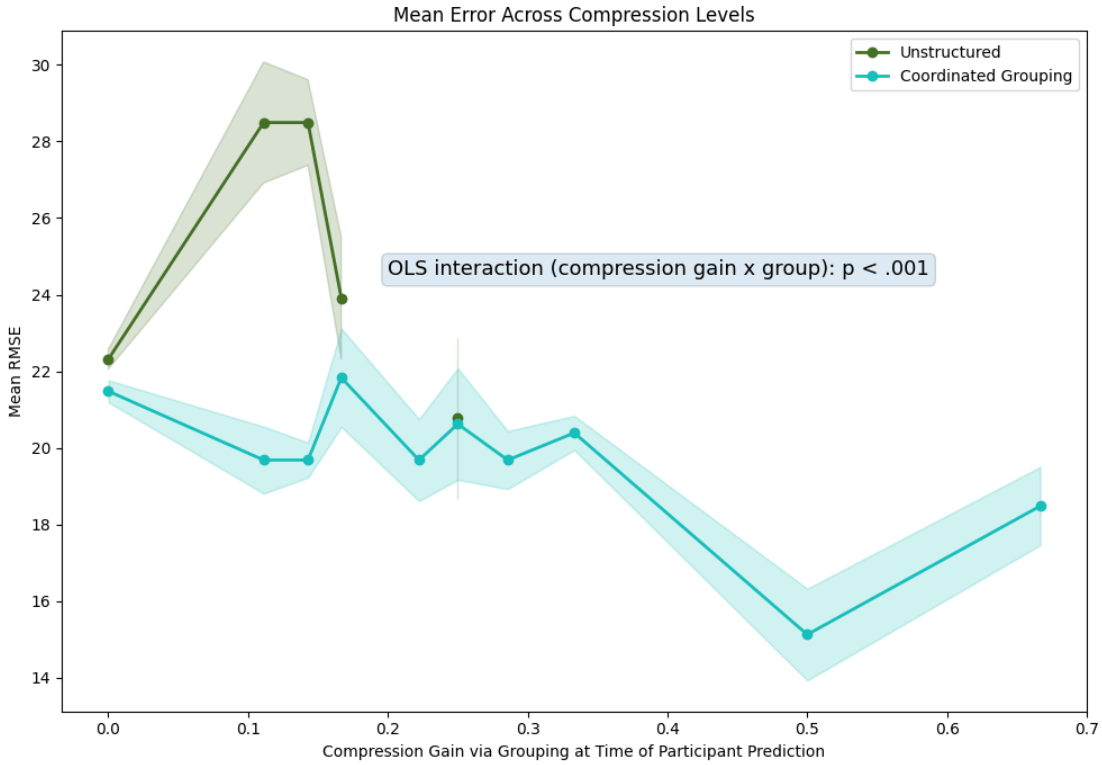

Figure 5.7: **Structural compression decreases prediction error (RMSE).** Each point represents prediction performance at a given level of compression gain, defined as $1 - \frac{N_{eff}}{N}$. Compression gain reflects the degree to which the multi-agent system can be represented using fewer effective units via pedestrian-inspired grouping. Results are shown separately for the independent-agent condition (baseline) and coordinated grouping (experimental) conditions. We fit an ordinary least squares regression model predicting RMSE from compression gain, group, and their interaction, with coefficient-specific p-values used to test main effects (conditional on the reference group or compression gain = 0) and the interaction term assessing whether the compression gain–RMSE relationship differed by group. The plot illustrates how prediction error varies as a function of structural compression, highlighting differences in how each condition utilizes or benefits from compressed representations of multi-agent motion, as well as the inability to achieve higher levels of compression without coordinated grouping.

5.5.4 Effective Number of Agents and Predictive Performance

To better quantify the impact of the use of coordinated grouping on prediction accuracy, we first examine the effective number of agents (N_{eff}), which serves as a measure of how many independent entities participants must track after accounting for grouping behavior induced by our algorithm. Lower values of N_{eff} indicate greater compression of agent behavior into higher-level group structures, thereby reducing perceived and computational task complexity.

Between conditions, the proposed algorithm consistently produced lower effective numbers of agents compared to the baseline ($p = 0.036$ overall), indicating that it successfully induced structured grouping and reduced the effective complexity of the multi-agent system. This compression effect was observed across rounds, though its behavioral impact varied depending on task difficulty.

To better capture relative compression across varying task sizes, we utilized two derived metrics: the compression ratio, and the compression gain, defined in Section 5.3.6. Together, these measures quantify the extent to which the system reduces effective complexity relative to the original agent set, enabling comparison across environments with different agent counts. As seen in Figure 5.6, without patterning, the organically occurring compression in most environments in nonexistent. Even with coordinated grouping, oftentimes the setting does not allow for compression gain due to the number of agents or their orientation. However, the use of dynamic pedestrian groups unlocks high compression in appropriate settings, which is not achievable with independent agents.

We fit an ordinary least squares regression model predicting RMSE from compression gain, group, and their interaction. The interaction term tested whether the relationship between compression gain and RMSE differed across groups. Main effects were interpreted conditionally: the effect of compression gain reflects its association with RMSE within the reference group, while the group effect reflects differences in RMSE at zero compression gain. Statistical significance was assessed using the coefficient-specific p-values from the model. Analysis of these metrics shows a consistent relationship between compression gain and predictive performance which can be seen in Figure 5.7: higher compression gain is associated with lower prediction error ($p < .001$). This

trend suggests that greater effective simplification of the multi-agent structure via patterning leads to improved human predictive accuracy. In other words, when the algorithm achieves stronger compression of the agent space, participants are better able to anticipate system behavior.

These findings reinforce the interpretation that performance improvements are driven not just by absolute reductions in effective agent count, but by the proportional degree of compression relative to task complexity. The dynamic utilization of pedestrian-inspired coordinated groups allows for greater compression of the environment. This in turn allows for greater abstraction of the environment and reduction in the number of entities to manage. As a result of this reduction, human prediction accuracy improves compared to the baseline.

5.6 Discussion

These findings indicate that the use of dynamic pedestrian-inspired grouping altered objective task performance without changing participants' subjective experience of the task. Across conditions, self-reported cognitive load, perceived predictability, and perceived understandability remained unchanged, suggesting that the intervention did not meaningfully affect how difficult or interpretable participants believed the system to be. At the same time, the intervention improved objective predictive accuracy based on RMSE under higher-complexity conditions. This pattern suggests that the benefits of coordinated grouping emerge selectively rather than uniformly. Rather than serving as a general enhancement to predictability, the intervention reaps benefits when task demands exceed a complexity threshold.

This conditionality is central to understanding the impact of pedestrian-inspired grouping. The use of coordinated grouping does not simplify the task through direct guidance or explicit assistance. Instead, it restructures the problem space. By enabling agents to dynamically form pedestrian-inspired groups, the system reduces the effective number of entities participants must reason about at any given time, replacing multiple agents with familiar pedestrian-like groups. This produces a form of observable compression, in which multiple agents can be represented as a smaller number of higher-order units. The results demonstrate that this compression was

measurably greater in the coordinated grouping condition, while high-compression regimes were absent in the baseline condition with classic planning methods. Importantly, these high-compression states were associated with improved predictive performance. Taken together, these findings suggest that the coordinated grouping does not directly improve prediction ability; rather, it changes the representational structure available to observers, enabling more efficient reasoning and accurate prediction in sufficiently complex environments.

This also clarifies why benefits emerged primarily in later rounds and at higher agent counts. Under low-complexity conditions, participants were capable of tracking the full system without relying on additional structure. In such settings, the cognitive demands of the task remained tractable, and the availability of compression provided little practical advantage. As complexity increased, however, maintaining accurate predictions through unstructured reasoning became progressively less effective. Under these circumstances, compressed representations became disproportionately useful, allowing participants to substitute detailed tracking of individual agents with reasoning over pedestrian-inspired groups. The intervention therefore appears to support performance not by reducing workload in absolute terms, but by enabling observers to operate at a more efficient level of abstraction when system complexity necessitates it.

These findings align with broader theories of cognitive offloading and abstraction in complex systems. In many domains, effective human reasoning is aided by the ability to reduce dimensionality by identifying meaningful structure within large information spaces. Coordinated grouping can be understood as an externalized form of such dimensionality reduction. This perspective situates this work within a wider body of work on human decision-making under complexity, in which abstraction serves as a critical mechanism for maintaining performance as task scale increases.

The implications of this work extend to emerging work in shared autonomy, explainable coordination, and adaptive interfaces for multi-agent environments. As autonomous systems become larger and more interdependent, the challenge of human oversight increasingly depends on how effectively system behavior can be represented and interpreted. The present results provide empirical evidence that structural organization, rather than merely additional information, can improve

human predictive ability in complex dynamic systems.

Several limitations should be considered when interpreting these results. redFirst, while the number of individual participants was relatively small ($n = 21$), the repeated-measures design (yielding 1,200 distinct prediction events) and the continuous, high-resolution nature of the RMSE trajectory data provided sufficient statistical power to detect the reported effects. Second, compression was inferred from system structure rather than directly measured as a cognitive process. While effective-agent metrics provide a principled estimate of representational complexity, they remain a proxy for participant mental models rather than a direct observation of cognition. Third, the self-report measures employed may not have been sufficiently sensitive to detect subtle internal shifts in reasoning strategy. Lastly, the observed performance benefits were context-dependent and not universal, emerging only under specific levels of complexity.

Future work should address these limitations by examining real-time adaptive interventions that respond dynamically to system complexity, as well as by directly measuring participant mental models through process-tracing, think-aloud protocols, or eye-tracking methodologies. It will also be valuable to explore whether training can increase participants' ability to exploit compressed representations, thereby extending the benefits of structural abstraction. Expanding this framework to other multi-agent and human-autonomy interaction domains will help determine its broader applicability. Additionally, because the present study was conducted in a controlled experimental context, and the social gaze aspect of pedestrian grouping was omitted, in-person or operational settings may produce different effects due to social, environmental, and temporal factors. These contexts should be investigated to better understand how structural interventions perform in realistic deployments.

5.7 Conclusion

Dynamically grouping agents and altering their collective behavior to follow recognizable pedestrian heuristics improved participant predictive accuracy in complex scenarios, demonstrating that performance gains can be achieved by organizing the system to mimic familiar structures.

Notably, these improvements emerged only in higher-complexity conditions, where the number of agents in the environment exceeded the threshold at which participants could accurately track them as independent entities. This suggests that the value of pedestrian-inspired grouping is not in simplifying the task in an absolute sense, but in reshaping how complexity is presented to the observer.

The findings reinforce the importance of structural compression in human interaction with dynamic multi-agent systems. By encouraging agents to move in patterned ways, the system reduced the number of effectively distinct elements participants needed to model. This enabled observers to form higher-order mental representations by tracking groups, flows, or collective behaviors rather than isolated individuals. Such compression appears to support prediction without necessarily reducing reported workload, reinforcing that objective performance and subjective effort can diverge.

This distinction has broader implications for the design of human-centered autonomous systems. Traditional approaches of human support in complex environments often focus on increasing transparency through more information, more detailed displays, explicit explanations, or improving human prediction. The results of this work suggest an alternative pathway: designing systems whose behavior is inherently predictable via familiar structures. Rather than requiring humans to process additional data, systems can be re-structured to reveal familiar structures that align with human perceptual and cognitive tendencies. In this way, predictability and interpretability become emergent properties of behavior rather than an afterthought.

Humans are especially adept at recognizing heuristically-guided movement and extracting rules from repeated exposure. Leveraging these tendencies allows complex systems to be more understandable to humans. Systems that move in ways that are perceptually coherent can support stronger human models of future behavior, without significantly altering behavior on the part of the agent.

As dynamic systems continue to grow in scale, autonomy, and interdependence, the challenge of human oversight will increasingly center on cognitive scalability. It is not sufficient for systems

to be accurate or efficient in isolation; they must also remain understandable to the people who monitor, collaborate with, or make decisions alongside them. The capacity to compress complexity into meaningful patterns is a key design principle for future autonomous and decision-support systems.

Ultimately, this work highlights a shift in perspective: effective human support in complex environments may depend less on reducing effort and more on enabling the right kind of abstraction. By structuring systems so that observers can reason about them at higher levels of organization, we can preserve performance even as complexity scales beyond the limits of direct attention. In that sense, the future of human-system collaboration may rest not on providing people with more information, but on making complexity intelligible through form.

5.8 Unifying Themes Across the Dissertation

This work establishes the role of patterns in a third major context: multi-agent human-robot teaming. Building on prior efforts to formalize patterns in both discrete planning and continuous interaction, this work extends the concept to settings where humans must coordinate with multiple robotic teammates simultaneously. In doing so, it demonstrates that patterns remain a useful organizing principle as team complexity increases, though not necessarily when multi-agent management is not sufficiently cognitively taxing. Rather than becoming less relevant in larger, more dynamic settings, patterns continue to provide structure that can support more predictable and coordinated robot behavior in highly complex environments.

Importantly, these results show that pattern-based approaches can still offer meaningful benefits in multi-agent teams, where the demands on human attention and decision-making are especially high. This suggests that patterns are not just a useful tool for simplifying isolated interactions, but a broader framework for designing effective human-robot collaboration in high complexity settings. Taken together, this work completes a progression that establishes patterns as a powerful mechanism for improving teaming at multiple levels of abstraction, setting the stage for a broader reflection on their implications, limitations, and future directions.

Chapter 6

Conclusion

6.1 Summary of Contributions and Key Takeaways

This dissertation makes several high-level contributions toward improving human–robot interaction through the lens of predictability and patterning. Across the included papers, a central contribution is demonstrating that patterns can be formally defined and systematically incorporated into robot decision-making. By moving beyond informal or ad hoc notions of predictability, this work provides a framework for reasoning about how structured behavior can be designed, analyzed, and deployed in human–robot teams in a variety of contexts.

The results consistently show that such formalized patterns are effective in practice. Robots that employ patterned behavior are not only more predictable, but also more understandable to human teammates. This improved predictability translates into stronger teaming performance: humans are better able to anticipate robot actions, coordinate their own behavior, and recover from deviations. In turn, these systems are perceived more favorably, suggesting that predictability and structure influence not just objective outcomes, but also subjective evaluations of the robot as a collaborator.

6.1.1 Establishing Patterning as a Viable Method for Teaming Improvement and Subtask Level Patterning

This dissertation shows that patterns can be formalized at the level of subtask planning, where structured, repeatable behaviors shape how agents act and interact over time. Introduc-

ing such patterning improves the predictability of robot behavior, allowing human partners to anticipate future actions more effectively. This increased predictability in turn supports greater understandability, as humans can form clearer internal representations of how the system operates without needing to track every individual agent or decision in detail.

These benefits extend beyond task performance to influence human perceptions of the robot itself. Patterned behavior improves how participants evaluate the robot, particularly in its role as a teammate, fostering a stronger sense of coordination and alignment at the subtask level. As a result, teams achieve better overall outcomes when structured patterns are present. Notably, while humans are able to successfully work with and leverage these patterns, they are often unable to explicitly articulate them, suggesting that the benefits of patterning operate at an implicit level of cognition rather than through conscious reasoning.

6.1.2 Balancing Patterning with Optimality and Patterning in Navigation

Patterns can also be formalized directly within the navigation layer, shaping how robots move through space in ways that are structured and predictable to human partners. Rather than pursuing purely optimal paths, navigation can be designed to balance efficiency with pattern consistency, introducing trajectories that may be slightly suboptimal but more predictable and easier to reason over. This trade-off allows robots to communicate intent implicitly through motion, supporting smoother coordination without requiring explicit signaling or explanation.

Incorporating patterning at this level also leads to stronger teaming outcomes. Human–robot teams benefit not only in objective performance metrics but also in how humans perceive the interaction. Patterned navigation produces more positive evaluations of the robot, particularly in terms of trust, fluency, and the robot's role as a collaborative partner. These findings suggest that embedding structure into low-level behavior can meaningfully influence both functional and social dimensions of teaming.

6.1.3 Multi-Agent Patterning

Extending patterning to multi-agent systems introduces additional complexity, as interactions are no longer limited to a single robot but emerge from the coordination of many agents. As the number of agents increases, the system becomes harder to track and predict, and the structure imposed by patterning plays a more critical role. By organizing agents into coherent, patterned behaviors, it becomes possible to reduce the system's effective complexity, enabling humans to reason about groups rather than individuals.

However, the benefits of patterning in this context are more conditional. Improvements in predictability emerge primarily when the degree of compression is sufficiently high—that is, when patterns meaningfully reduce the number of independent elements a human must track. At lower levels of compression, the added structure does not translate into measurable gains, as the differences between patterned movement and independent movement are not significant. Additionally, these improvements do not appear to extend to subjective experience: participants do not report increased ease, understanding, or reduced cognitive load, even when their predictive performance improves. This suggests that patterning operates by restructuring the task at a more intuitive cognitive level, as also seen with PACT, rather than altering how difficult the task feels.

6.1.4 Cross-Cutting Insights

An additional finding across this work is that increased behavioral complexity does not necessarily correspond to increased perceived intelligence. More complex or less structured behaviors can, in many cases, hinder understanding and reduce effective coordination. In contrast, simpler, well-structured patterns often lead to stronger impressions of competence and reliability. This highlights an important design implication: optimizing for human interpretability may be more valuable than maximizing behavioral sophistication in isolation.

Finally, this dissertation provides evidence that patterned behavior can help stabilize human mental models of robot teammates over time. By presenting consistent and predictable structures,

robots enable users to form durable expectations that persist across interactions. As shown in the work with multiple agents, patterns can also be useful in certain group contexts. Additionally, the greater the compression of the agents via patterning, the more said patterning helps. More specifically, patterning increases accuracy the more it reduces environmental complexity. Together, these findings reinforce the broader claim that predictability, achieved through formalized patterns, plays a critical role in shaping both the effectiveness and the experience of human–robot teaming.

6.2 Implications for Future Work

Predictability will continue to be a key design necessity in the future of human–robot interaction, particularly as systems are deployed in increasingly complex, real-world environments. The results of this thesis reinforce that predictability is not merely a byproduct of good system design, but a primary factor shaping team performance and coordination quality. As HRI systems scale to include more agents, longer time horizons, and less structured tasks, ensuring that robot behavior remains interpretable and predictable will become even more critical.

The use of structured behavioral patterns represents a practical and extensible mechanism for achieving this predictability. Patterns provide a compact way to encode and communicate behavior, allowing humans to quickly learn and adapt to robot actions without requiring detailed knowledge of the underlying algorithms. Future systems can build on this idea by developing richer pattern libraries, adapting patterns dynamically to context, or enabling robots to explicitly signal which pattern they are following. Such directions could further reduce cognitive load while maintaining flexibility in coordination.

An important implication for future work is the role of mental model stability. While this dissertation focuses on establishing and leveraging patterns, maintaining consistency in how those patterns are presented and executed over time may be just as important. If human teammates can form stable, reliable expectations about robot behavior, they are better equipped to generalize across tasks and recover from unexpected situations. Designing for this stability; particularly in adaptive or learning systems, presents a key challenge, as improvements in performance must be

balanced against the risk of disrupting established expectations.

A second promising direction is the role of patterning in trust calibration and intent alignment. While predictable behaviors can improve coordination, excessive regularity may also lead human teammates to over-trust autonomous systems or make incorrect assumptions about their capabilities. Future work could investigate how robot teams can communicate not only what they are doing, but also the confidence, uncertainty, or rationale underlying those behaviors. This raises important questions about how patterns should adapt when robot objectives diverge from human expectations, and how transparency mechanisms can help preserve calibrated trust without sacrificing efficiency. Exploring these dynamics may enable robot teams that are both predictable and appropriately interpretable, particularly in high-stakes or rapidly changing environments.

Another important avenue for future research involves the measurement and modeling of cognitive load in human–robot teaming. This dissertation primarily evaluates outcomes through task performance and subjective perception metrics, but future systems may benefit from more direct, continuous estimates of cognitive effort. Incorporating physiological signals, behavioral indicators, or adaptive workload models could provide deeper insight into when structural compression and patterning are most beneficial. Such measures may also help explain the nonlinear effects observed in multi-agent settings, where interventions that improve predictability do not always translate to reduced subjective burden. Understanding these relationships could support the development of adaptive teaming strategies that dynamically balance efficiency, predictability, and human cognitive capacity in real time.

More broadly, these directions point toward a future in which human–robot teams are designed not solely around task optimization, but around the long-term maintenance of shared understanding between humans and autonomous systems. As robot teams become larger, more adaptive, and more autonomous, mechanisms that support stable mental models, calibrated trust, and manageable cognitive demands will likely become increasingly central to effective collaboration.

Overall, these considerations point toward a unifying objective for future HRI research: to design systems that not only act effectively, but do so in ways that remain predictable, consis-

tent, and understandable over time. Emphasizing predictability, leveraging structured patterns, and preserving mental model stability together provide a foundation for more robust and scalable human–robot teaming.

Bibliography

[1] Charles F. Abel. Heuristics and problem solving. New Directions for Teaching and Learning, 2003(95):53–58, 2003.

[2] Stéphane Airiau, Sandip Sen, and Daniel Villatoro. Emergence of conventions through social learning: Heterogeneous learners in complex networks. Autonomous Agents and Multi-Agent Systems, 28(5):779–804, 2014.

[3] Fatima M. Albar and Antonie J. Jetter. Heuristics in decision making. PICMET: Portland International Center for Management of Engineering and Technology, Proceedings, pages 578–584, 2009.

[4] Stéphane Aroca-Ouellette, Miguel Aroca-Ouellette, Katharina von der Wense, and Alessandro Roncone. Implicitly aligning humans and autonomous agents through shared task abstractions. In Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25. International Joint Conferences on Artificial Intelligence Organization, 2025. Main Track.

[5] Stéphane Aroca-Ouellette, Miguel Aroca-Ouellette, Upasana Biswas, Katharina Kann, and Alessandro Roncone. Hierarchical reinforcement learning for ad hoc teaming. In Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems, pages 2337–2339, 2023.

[6] Franziska Babel, Johannes Kraus, and Martin Baumann. Findings From A Qualitative Field Study with An Autonomous Robot in Public : Exploration of User Reactions and Conflicts. International Journal of Social Robotics, 14:1625–1655, 2022.

[7] Daniel Belanche, Luis V Casaló, Jeroen Schepers, and Carlos Flavián. Examining the effects of robots' physical appearance, warmth, and competence in frontline services: The humanness-value-loyalty model. Psychology & Marketing, 38(12):2357–2376, 2021.

[8] Debjit Bhowmick, Stephan Winter, Mark Stevenson, and Peter Vortisch. The impact of urban road network morphology on pedestrian wayfinding behaviour. Journal of Spatial Information Science, 12 2020.

[9] Serena Booth, Sanjana Sharma, Sarah Chung, Julie Shah, and Elena L Glassman. Revisiting Human-Robot Teaching and Learning Through the Lens of Human Concept Learning. In HRI '22: Proceedings of the 2022 ACM/IEEE International Conference on Human-Robot Interaction, pages 147–156, 2022.

[10] Florian Brachten, Felix Brünker, Nicholas R. J. Frick, Björn Ross, and Stefan Stieglitz. On the ability of virtual agents to decrease cognitive load: an experimental study. Information Systems and e-Business Management, 18(2):187–207, 2020.

[11] Connor Brooks and Daniel Szafir. Building second-order mental models for human-robot interaction. In Proceedings of the AAAI Fall Symposium Series: AI for Service Robots in Human Environments (AI-HRI '19), 2019.

[12] Tad Brunyé, Shaina Martis, and Holly Taylor. Cognitive load during route selection increases reliance on spatial heuristics. The Quarterly Journal of Experimental Psychology, 71:1–38, 03 2017.

[13] Tad Brunyé, Shaina Martis, and Holly Taylor. Cognitive load during route selection increases reliance on spatial heuristics. The Quarterly Journal of Experimental Psychology, 71:1–38, 03 2017.

[14] Erdem Bıyık, Anusha Lalitha, Rajarshi Saha, Andrea Goldsmith, and Dorsa Sadigh. Partner-Aware Algorithms in Decentralized Cooperative Bandit Teams. 2021.

[15] Fanta Camara, Nicola Bellotto, Serhan Cosar, Florian Weber, Dimitris Nathanael, Matthias Althoff, Jingyuan Wu, Johannes Ruenz, André Dietrich, Gustav Markkula, Anna Schieben, Fabio Tango, Natasha Merat, and Charles Fox. Pedestrian models for autonomous driving part ii: High-level models of human behavior. IEEE Transactions on Intelligent Transportation Systems, 22(9):5453–5472, 2021.

[16] Colleen M Carpinella, Alisa B Wyman, Michael A Perez, and Steven J Stroessner. The robotic social attributes scale (rosas) development and validation. In Proceedings of the 2017 ACM/IEEE International Conference on human-robot interaction, pages 254–262, 2017.

[17] Micah Carroll, Rohin Shah, Mark K. Ho, Thomas L. Griffiths, Sanjit A. Seshia, Pieter Abbeel, and Anca Dragan. On the Utility of Learning about Humans for Human-AI Coordination. Curran Associates Inc., Red Hook, NY, USA, 2019.

[18] Wesley P. Chan, Geoffrey Hanks, Maram Sakr, Haomiao Zhang, Tiger Zuo, H.F. Machiel Van der Loos, and Elizabeth Croft. Design and Evaluation of an Augmented Reality Head-Mounted Display Interface for Human Robot Teams Collaborating in Physically Shared Manufacturing Tasks. ACM Transactions on Human-Robot Interaction, pages 1–19, 2022.

[19] Christine T. Chang and Bradley Hayes. A survey of augmented reality for human–robot collaboration. Machines, 12(8), 2024.

[20] Rui Chen, Alvin Shek, and Changliu Liu. Learn from human teams: a probabilistic solution to real-time collaborative robot handling with dynamic gesture commands. CoRR, abs/2112.06020, 2021.

[21] Xiaobei Chen, Fanjue Liu, and Luling Huang. What did you hear and what did you see? understanding the transparency of facial recognition and speech recognition systems during human–robot interaction. New Media Society, 27:5776–5802, 06 2024.

[22] Nicolette Peterson Christopher Curry, Ruixuan Li and Thomas A. Stoffregen. Cybersickness in virtual reality head-mounted displays: Examining the influence of sex differences and

vehicle control. International Journal of Human–Computer Interaction, 36(12):1161–1167, 2020.

[23] Mark Colley, Christian Bräuner, Mirjam Lanzer, Marcel Walch, Martin Baumann, and Enrico Rukzio. Effect of visualization of pedestrian intention recognition on trust and cognitive load. In 12th International Conference on Automotive User Interfaces and Interactive Vehicular Applications, AutomotiveUI '20, page 181–191, New York, NY, USA, 2020. Association for Computing Machinery.

[24] L. J. Cronbach. Coefficient alpha and the internal structure of tests. Psychometrika, 16:297–334, 1951.

[25] Felipe Cucker and Steve Smale. Emergent behavior in flocks. IEEE Transactions on Automatic Control, 52(5):852–862, 2007.

[26] Hongjun Cui, Jinping Xie, Minqing Zhu, Xiaoyong Tian, and Ce Wan. Virus transmission risk of college students in railway station during post-covid-19 era: Combining the social force model and the virus transmission model. Physica A: Statistical Mechanics and its Applications, 608:128284, 2022.

[27] Sylvain Daronnat, Leif Azzopardi, Martin Halvey, and Mateusz Dubiel. Inferring Trust From Users' Behaviours; Agents' Predictability Positively Affects Trust, Task Performance and Cognitive Load in Human-Agent Real-Time Collaboration. Frontiers in Robotics and AI, 8(July):1–14, 2021.

[28] Maryam Banitalebi Dehkordi, Reda Mansy, Abolfazl Zaraki, Arpit Singh, and Rossitza Setchi. Explainability in human-robot teaming. Procedia Computer Science, 192:3487–3496, 2021. Knowledge-Based and Intelligent Information Engineering Systems: Proceedings of the 25th International Conference KES2021.

[29] Ning Ding, Yu Zhu, Xinyan Liu, Dapeng Dong, and Yang Wang. A modified social force model for crowd evacuation considering collision predicting behaviors. Applied Mathematics and Computation, 466:128448, 2024.

[30] Anca D. Dragan, Shira Bauman, Jodi Forlizzi, and Siddhartha S. Srinivasa. Effects of robot motion on human-robot collaboration. In Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction, HRI '15, page 51–58, New York, NY, USA, 2015. Association for Computing Machinery.

[31] Anca D. Dragan, Kenton C.T. Lee, and Siddhartha S. Srinivasa. Legibility and predictability of robot motion. In Proceedings of the 8th ACM/IEEE International Conference on Human-Robot Interaction, HRI '13, page 301–308. IEEE Press, 2013.

[32] Daniel S Drew. Multi-agent systems for search and rescue applications. Current Robotics Reports, 2(2):189–200, 2021.

[33] Katherine Driggs-Campbell and Ruzena Bajcsy. Communicating intent on the road through human-inspired control schemes. In 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pages 3042–3047. IEEE, 2016.

[34] Ricardo Duarte, Duarte Araújo, Vanda Correia, and Keith Davids. Sports teams as superorganisms: Implications of sociobiological models of behaviour for research and practice in team sports performance analysis. Sports medicine, 42(8):633–642, 2012.

[35] Xiaocong Fan and John Yen. Modeling cognitive loads for evolving shared mental models in human–agent collaboration. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 41(2):354–367, 2011.

[36] Jaime Fernández Fisac, Chang Liu, Jessica B. Hamrick, S. Shankar Sastry, J. Karl Hedrick, Thomas L. Griffiths, and Anca D. Dragan. Generating plans that predict themselves. ArXiv, abs/1802.05250, 2018.

[37] Scott Forer, Santosh Balajee Banisetty, Logan Yliniemi, Monica Nicolescu, and David Feil-Seifer. Socially-aware navigation using non-linear multi-objective optimization. In IEEE/RSJ International Conference on Intelligent Robots and Systems, Madrid, Spain, October 2018.

[38] Rinat Galin and Roman Meshcheryakov. Human-Robot Interaction Efficiency and Human-Robot Collaboration, pages 55–63. 01 2020.

[39] Ali Ghadirzadeh, Xi Chen, Wenjie Yin, Zhengrong Yi, Marten Bjorkman, and Danica Kragic. Human-Centered Collaborative Robots with Deep Reinforcement Learning. IEEE Robotics and Automation Letters, 6(2):566–571, 2021.

[40] Gerd Gigerenzer. Why heuristics work. Perspectives on Psychological Science, 3(1):20–29, 2008. PMID: 26158666.

[41] Ji Han, Gopika Ajaykumar, Ze Li, and Chien Ming Huang. Structuring Human-Robot Interactions via Interaction Conventions. 29th IEEE International Conference on Robot and Human Interactive Communication, RO-MAN 2020, pages 341–348, 2020.

[42] Marc Hanheide, Annika Peters, and Nicola Bellotto. Analysis of human-robot spatial behaviour applying a qualitative trajectory calculus. Proceedings - IEEE International Workshop on Robot and Human Interactive Communication, (September):689–694, 2012.

[43] Glenda Hannibal and Felix Lindner. Towards a Questions-Centered Approach to Explainable Human-Robot Interaction. 01 2023.

[44] Sandra G. Hart and Lowell E. Staveland. Development of nasa-tlx (task load index): Results of empirical and theoretical research. In Peter A. Hancock and Najmedin Meshkati, editors, Human Mental Workload, volume 52 of Advances in Psychology, pages 139–183. North-Holland, 1988.

[45] Sergiu Hart. Adaptive heuristics. Econometrica, 73(5):1401–1430, 2005.

[46] Robert D Hawkins, Michael Franke, Michael C Frank, Adele E Goldberg, Kenny Smith, Thomas L Griffiths, and Noah D Goodman. From partners to populations: A hierarchical bayesian account of coordination and convention. Psychological Review, 130(4):977, 2023.

[47] Bradley Hayes and Brian Scassellati. Effective robot teammate behaviors for supporting sequential manipulation tasks. In 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pages 6374–6380. IEEE, 2015.

[48] Dirk Helbing and Péter Molnár. Social force model for pedestrian dynamics. Phys. Rev. E, 51:4282–4286, May 1995.

[49] Nicholas Hetherington, Elizabeth Croft, and H.F. Loos. Hey robot, which way are you going nonverbal motion legibility cues for human-robot interaction. IEEE Robotics and Automation Letters, PP:1–1, 03 2021.

[50] Guy Hoffman. Evaluating Fluency in Human-Robot Collaboration. IEEE Transactions on Human-Machine Systems, 49(3):209–218, 2019.

[51] Hengyuan Hu, Adam Lerer, Brandon Cui, Luis Pineda, Noam Brown, and Jakob Foerster. Off-belief learning. In International Conference on Machine Learning, pages 4369–4379. PMLR, 2021.

[52] Hengyuan Hu, Adam Lerer, Alex Peysakhovich, and Jakob Foerster. "Other-play" for zero-shot coordination. In Hal Daumé III and Aarti Singh, editors, Proceedings of the 37th International Conference on Machine Learning, volume 119 of Proceedings of Machine Learning Research, pages 4399–4410. PMLR, 13–18 Jul 2020.

[53] Yuto Imamura, Kazunori Terada, and Hideyuki Takahashi. Effects of behavioral complexity on intention attribution to robots. In Proceedings of the 3rd International Conference on Human-Agent Interaction, pages 65–72, 2015.

[54] Abhinav Jain, Daphne Chen, Dhruva Bansal, Sam Scheele, Mayank Kishore, Hritik Sapra, David Kent, Harish Ravichandar, and Sonia Chernova. Anticipatory human-robot collaboration via multi-objective trajectory optimization. In 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pages 11052–11057, 2020.

[55] Chao Jiang, Zhen Ni, Yi Guo, and Haibo He. Learning human–robot interaction for robot-assisted pedestrian flow optimization. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 49(4):797–813, 2019.

[56] Catholijn M. Jonker, M. Birna van Riemsdijk, and Bas Vermeulen. Shared mental models. In Marina De Vos, Nicoletta Fornara, Jeremy V. Pitt, and George Vouros, editors, Coordination, Organizations, Institutions, and Norms in Agent Systems VI, pages 132–151, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg.

[57] Yuka Kato, Yuka Nagano, and Haruka Yokoyama. A pedestrian model in human-robot coexisting environment for mobile robot navigation. In 2017 IEEE/SICE International Symposium on System Integration (SII), pages 992–997, 2017.

[58] Shauharda Khadka, Somdeb Majumdar, Tarek Nassar, Zach Dwiel, Evren Turner, Santiago Miret, Yinyin Liu, and Kagan Turner. Collaborative evolutionary reinforcement learning. 36th International Conference on Machine Learning, ICML 2019, 2019-June:5816–5827, 2019.

[59] Andreas Kolling, Phillip Walker, Nilanjan Chakraborty, Katia Sycara, and Michael Lewis. Human interaction with robot swarms: A survey. IEEE Transactions on Human-Machine Systems, 46(1):9–26, 2015.

[60] Georgios Kostopoulos, Gregory Davrazos, and Sotiris Kotsiantis. Explainable artificial intelligence-based decision support systems: A recent review. Electronics, 13(14):2842, 2024.

[61] W.H. Kruskal and W.A. Wallis. Use of ranks in one-criterion analysis of variance. Journal of the American Statistical Association, 47(260):583–621, 1952.

[62] Anton Kuznietsov, Balint Gyevnar, Cheng Wang, Steven Peters, and Stefano V Albrecht. Explainable ai for safe and trustworthy autonomous driving: A systematic review. IEEE Transactions on Intelligent Transportation Systems, 2024.

[63] Minae Kwon, Erdem Biyik, Aditi Talati, Karan Bhasin, Dylan P. Losey, and Dorsa Sadigh. When humans aren't optimal: Robots that collaborate with risk-aware humans. ACM/IEEE International Conference on Human-Robot Interaction, pages 43–52, 2020.

[64] Minae Kwon, Sandy H. Huang, and Anca D. Dragan. Expressing robot incapability. In Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction, HRI '18, page 87–95, New York, NY, USA, 2018. Association for Computing Machinery.

[65] Minae Kwon, Malte F. Jung, and Ross A. Knepper. Human expectations of social robots. ACM/IEEE International Conference on Human-Robot Interaction, 2016-April:463–464, 2016.

[66] Steven LaValle. Rapidly-exploring random trees: A new tool for path planning. Research Report 9811, 1998.

[67] Adam Lerer and Alexander Peysakhovich. Learning existing social conventions via observationally augmented self-play. AIES 2019 - Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, pages 107–114, 2019.

[68] Fei Li, L. Phillip Wang, Xiaoming Shen, and Joe Z. Tsien. Balanced dopamine is critical for pattern completion during associative memory recall. PLoS ONE, 5(10), 2010.

[69] Zhengping Li, Cheng Hwee Sim, and Malcolm Yoke Hean Low. A survey of emergent behavior and its impacts in agent-based systems. In 2006 4th IEEE international conference on industrial informatics, pages 1295–1300. IEEE, 2006.

[70] Jiaming Liu, Hui Zhang, Ning Ding, and Yuntao Li. A modified social force model for sudden attack evacuation based on yerkes–dodson law and the tendency toward low risk areas. Physica A: Statistical Mechanics and its Applications, 633:129403, 2024.

[71] Clare Lohrmann, Maria Stull, Alessandro Roncone, and Bradley Hayes. Generating pattern-based conventions for predictable planning in human-robot collaboration. ACM Transactions on Human-Robot Interaction, mar 2024.

[72] Clare Lohrmann, Maria P. Stull, Breanne Crockett, Calvin Ferraro, Ethan Berg, Alessandro Roncone, and Bradley Hayes. Thinking in patterns: Sacrificing performance for predictability enhances human-ai teams. Science Robotics, 2026.

[73] Laura Londoño, Adrian Röfer, Tim Welschehold, and Abhinav Valada. Doing Right by Not Doing Wrong in Human-Robot Collaboration. 2022.

[74] Luca Longo and Stephen Barrett. Cognitive effort for multi-agent systems. In Yiyu Yao, Ron Sun, Tomaso Poggio, Jiming Liu, Ning Zhong, and Jimmy Huang, editors, Brain Informatics, pages 55–66, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg.

[75] Ramya Mandava, Sai Srinivas Vellela, Shobana Gorintla, Lavanya Dalavai, Nallapu Malathi, and Koya Haritha. Evaluating the impact of explainable ai on user trust in financial decision-support systems. In 2025 International Conference on Computational Robotics, Testing and Engineering Evaluation (ICCRTEE), pages 1–6. IEEE, 2025.

[76] Olivier Mangin, Alessandro Roncone, and Brian Scassellati. How to be helpful? supportive behaviors and personalization for human-robot collaboration. Frontiers in Robotics and AI, 8, 2022.

[77] Michelle A Marks, Mark J Sabella, C Shawn Burke, and Stephen J Zaccaro. The impact of cross-training on team effectiveness. Journal of Applied Psychology, 87(1):3, 2002.

[78] Barnaby Marsh. Heuristics as social tools. New Ideas in Psychology, 20(1):49–57, 2002.

[79] John E. Mathieu, Gerald F. Goodwin, Tonia S. Heffner, Eduardo Salas, and Janis A. Cannon-Bowers. The influence of shared mental models on team process and performance. Journal of Applied Psychology, 85(2):273–283, 2000.

[80] Mark P. Mattson. Superior pattern processing is the essence of the evolved human brain. Frontiers in Neuroscience, 8(8 AUG):1–17, 2014.

[81] Jennifer Misyak, Takao Noguchi, and Nick Chater. Instantaneous conventions: The emergence of flexible communicative signals. Psychological science, 27(12):1550–1561, 2016.

[82] Shuwa Miura, Andrew L Cohen, and Shlomo Zilberstein. Maximizing legibility in stochastic environments. In 2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN), pages 1053–1059. IEEE, 2021.

[83] Susan Mohammed, Lori Ferzandi, and Katherine Hamilton. Metaphor no more: A 15-year review of the team mental model construct. Journal of management, 36(4):876–910, 2010.

[84] Luis Yoichi Morales Saiki, Satoru Satake, Rajibul Huq, Dylan Glas, Takayuki Kanda, and Norihiro Hagita. How do people walk side-by-side? using a computational model of human behavior for a social robot. In Proceedings of the Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction, HRI '12, page 301–308, New York, NY, USA, 2012. Association for Computing Machinery.

[85] Shabnam Mousavi and Gerd Gigerenzer. Risk, uncertainty, and heuristics. Journal of Business Research, 67(8):1671–1678, 2014.

[86] Sara Moussawi, Marios Koufaris, and Raquel Benbunan-Fich. How perceptions of intelligence and anthropomorphism affect adoption of personal intelligent agents. Electronic Markets, 31(2):343–364, 2021.

[87] Mehdi Moussaïd, Dirk Helbing, and Guy Theraulaz. How simple rules determine pedestrian behavior and crowd disasters. Proceedings of the National Academy of Sciences of the United States of America, 108:6884–8, 04 2011.

[88] Mehdi Moussaïd, Niriaska Perozo, Simon Garnier, Dirk Helbing, and Guy Theraulaz. The walking behaviour of pedestrian social groups and its impact on crowd dynamics. PloS one, 5:e10047, 04 2010.

[89] Manisha Natarajan, Esmaeil Seraj, Batuhan Altundas, Rohan Paleja, Sean Ye, Letian Chen, Reed Jensen, Kimberlee Chang, and Matthew Gombolay. Human-robot teaming: Grand challenges. Current Robotics Reports, 4:1–20, 08 2023.

[90] Stefanos Nikolaidis and Julie A. Shah. Human-Robot Teaming using Shared Mental Models. IEEE/ACM International Conference on Human-Robot Interaction, Workshop on Human-Agent-Robot Teamwork (2012), 17(6):1098–1106, 2012.

[91] Tatsuya Nomura, Tomohiro Suzuki, Takayuki Kanda, and Kensuke Kato. Measurement of negative attitudes toward robots. Interaction Studies, 7(3):437–454, 2006.

[92] Amit Kumar Pandey and Rachid Alami. A framework for adapting social conventions in a mobile robot motion in human-centered environment. 2009 International Conference on Advanced Robotics, ICAR 2009, 2009.

[93] Max Pascher, Uwe Gruenefeld, Stefan Schneegass, and Jens Gerken. How to communicate robot motion intent: A scoping review. Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, 2023.

[94] 2025.

[95] Bethany Rittle-Johnson, Emily R. Fyfe, Laura E. McLean, and Katherine L. McEldoon. Emerging understanding of patterning in 4-year-olds. Journal of Cognition and Development, 14(3):376–396, 2013.

[96] Alessandra Rossi, Fernando Garcia, Arturo Cruz Maya, Kerstin Dautenhahn, Kheng Lee Koay, Michael L. Walters, and Amit K. Pandey. Investigating the Effects of Social Interactive Behaviours of a Robot on People's Trust During a Navigation Task. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 11649 LNAI(July):349–361, 2019.

[97] Shahabedin Sagheb, Soham Gandhi, and Dylan P. Losey. Should collaborative robots be transparent? International Journal of Social Robotics, 17(6):937–953, Jun 2025.

[98] Basak Sakcak and Luca Bascetta. Safe Motion Planning for a Mobile Robot Navigating in Environments Shared with Humans. 2022.

[99] Kristin E. Schaefer. Measuring Trust in Human Robot Interactions: Development of the "Trust Perception Scale-HRI", pages 191–218. Springer US, Boston, MA, 2016.

[100] Svenja Y. Schött, Rifat Mehreen Amin, and Andreas Butz. A literature survey of how to convey transparency in co-located human–robot interaction. Multimodal Technologies and Interaction, 7(3), 2023.

[101] Francesco Semeraro, Alexander Griffiths, and Angelo Cangelosi. Human–robot collaboration and machine learning: A systematic review of recent research. Robotics and Computer-Integrated Manufacturing, 79:102432, 2023.

[102] Andy Shih, Arjun Sawhney, Jovana Kondic, Stefano Ermon, and Dorsa Sadigh. On the critical role of conventions in adaptive human-ai collaboration. In Proceedings of the 9th International Conference on Learning Representations (ICLR), may 2021.

[103] Masahiro Shiomi, Francesco Zanlungo, Kotaro Hayashi, and Takayuki Kanda. Towards a socially acceptable collision avoidance for a mobile robot navigating among pedestrians using a pedestrian model. International Journal of Social Robotics, 6:443–455, 08 2014.

[104] Jamie Snape, Jur van den Berg, Stephen Guy, and Dinesh Manocha. Smooth and collision-free navigation for multiple robots under differential-drive constraints. pages 4584–4589, 10 2010.

[105] David Sobrín-Hidalgo, Ángel Guerrero-Higueras, and Vicente Matellán. Generating explanations for autonomous robots: A systematic review. IEEE Access, PP:1–1, 01 2025.

[106] DJ Strouse, Kevin McKee, Matt Botvinick, Edward Hughes, and Richard Everett. Collaborating with humans without human data. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, Advances in Neural Information Processing Systems, volume 34, pages 14502–14515. Curran Associates, Inc., 2021.

[107] Richard S Sutton, Andrew G Barto, et al. Reinforcement learning: An introduction, volume 1. MIT press Cambridge, 1998.

[108] Aaquib Tabrez, Matthew B. Luebbers, and Bradley Hayes. A Survey of Mental Modeling Techniques in Human–Robot Teaming. Current Robotics Reports, 1(4):259–267, 2020.

[109] Kartik Talamadupula, J. Benton, Subbarao Kambhampati, Paul Schermerhorn, and Matthias Scheutz. Planning for human-robot teaming in open worlds. ACM Transactions on Intelligent Systems and Technology, 1:14:1–14:24, 2010.

[110] Matthias Sebastian Treder. Behind the looking-glass: A review on human symmetry perception. Symmetry, 2(3):1510–1543, 2010.

[111] Mycal Tucker, Yilun Zhou, and Julie Shah. Latent space alignment using adversarially guided self-play. International Journal of Human–Computer Interaction, 0(0):1–19, 2022.

[112] Yakup Turgut and Cafer Erhan Bozdag. Modeling pedestrian group behavior in crowd evacuations. Fire and Materials, 46(2):420–442, 2022.

[113] Inara Tusseyeva, Anara Sandygulova, and Matteo Rubagotti. Perceived intelligence in human-robot interaction–a review. IEEE Access, 2024.

[114] Amos Tversky and Daniel Kahneman. Judgment under uncertainty: Heuristics and biases. Science, 185(4157):1124–1131, 1974.

[115] Amos Tversky and Daniel Kahneman. Advances in prospect theory: Cumulative representation of uncertainty. Journal of Risk and uncertainty, 5(4):297–323, 1992.

[116] Sjir Uitdewilligen, Mary J Waller, and Adrian H Pitariu. Mental model updating and team adaptation. Small Group Research, 44(2):127–158, 2013.

[117] Jur van den Berg, Stephen Guy, Ming Lin, and Dinesh Manocha. Reciprocal n-Body Collision Avoidance, volume 70, pages 3–19. 04 2011.

[118] Vasiliki Vouloutsi, Klaudia Grechuta, Stéphane Lallée, and Paul FMJ Verschure. The influence of behavioral complexity on robot perception. In Conference on Biomimetic and Biohybrid Systems, pages 332–343. Springer, 2014.

[119] Johan Wagemans. Characteristics and models of human symmetry detection. Trends in Cognitive Sciences, 1:346–352, 1997.

[120] Zhiqiang Wan, Chao Jiang, Muhammad Fahad, Zhen Ni, Yi Guo, and Haibo He. Robot-assisted pedestrian regulation based on deep reinforcement learning. IEEE Transactions on Cybernetics, 50(4):1669–1682, 2020.

[121] William Warren. Collective motion in human crowds. Current Directions in Psychological Science, 27:096372141774674, 07 2018.

[122] J. R. Wilson and A. Rutherford. Mental models: Theory and application in human factors. Human Factors, 31(6):617–634, 1989.

[123] Qian Xu, Wenzhao Xie, Bolin Liao, Chao Hu, Lu Qin, Zhengzijin Yang, Huan Xiong, Yi Lyu, Yue Zhou, and Aijing Luo. Interpretability of clinical decision support systems based on artificial intelligence from technological and medical perspective: A systematic review. Journal of healthcare engineering, 2023(1):9919269, 2023.

[124] Fangkai Yang and Christopher Peters. Social-aware navigation in crowds with static and dynamic groups. 09 2019.

[125] John Yen, Xiaocong Fan, Shuang Sun, Rui Wang, Cong Chen, Kaivan Kamali, and Richard a Volz. Implementing Shared Mental Models for Collaborative Teamwork. The Workshop on Collaboration Agents: Autonomous Agents for Collaborative Environments in the EEE/WIC Intelligent Agent Technology Conference, Halifax, Canada, 2003.

[126] Bingqing Zhang, Javad Amirian, Harry Eberle, Julien Pettre, Catherine Holloway, and Tom Carlson. From hri to cri: Crowd robot interaction—understanding the effect of robots on crowd motion. 06 2021.

Appendix A

PACT Appendix

A.1 PACT Algorithm

Algorithm 1 Best Pattern Selection

Input: Set of tasks T, Set of Patterns P
Output: The pattern(s) best suited for T

$minScore \leftarrow \infty$
$bestPatterns \leftarrow \emptyset$
for $p \in P$ **do**
  $score \leftarrow 0$
  for $i \in 1 \leq i \leq |T|$ **do**
    $S_p \leftarrow$ every allowable sequence of length $i-1$ using p
    $T_{i,p} \leftarrow []$
    for $s \in S_p$ **do**
      $t_s \leftarrow$ all allowable next tasks after completing s, under pattern p
      $T_{i,p}.extend(t_s)$
    end for
    $firstTerm = H(T_{i,p})$ // Calculate entropy
    $P_{i,\text{shared}} \leftarrow \{\}$ // Patterns sharing candidate seqs with p
    $T_{i,\text{shared}} \leftarrow []$
    for $q \in P$ **do**
      $S_q \leftarrow$ every allowable sequence of length $i-1$ using q
      $S_q = S_q \cap S_p$ // Only sequences that also follow p
      if $|S_q| > 0$ **then**
        $P_{i,\text{shared}} \leftarrow P_{i,\text{shared}} \cup \{q\}$
        for $s \in S_q$ **do**
          $t_s \leftarrow$ all allowable next tasks after completing s, under pattern q
          $T_{i,\text{shared}}.extend(t_s)$
        end for
      end if
    end for
    $discount = \frac{|P_{i,\text{shared}}|-1}{|P|}$
    $secondTerm = discount * H(T_{i,\text{shared}})$
    $score \leftarrow score + firstTerm + secondTerm$
  end for
  if $score = minScore$ **then**
    $bestPatterns \leftarrow bestPatterns \cup \{p\}$
  else if $score < minScore$ **then**
    $minScore = score$
    $bestPatterns \leftarrow \{p\}$
  end if
end for
return $bestPatterns$

A.2 PACT Survey Questions

Listed p-values are of the form (conventions/median, conventions/optimal, median/optimal).

A.2.1 Pre-Activity Survey

A.2.1.1 Experience with Robots

Questions in this section were either multiple choice, or select all that apply. Options for each question are listed below the question.

- Have you ever watched a movie or television show that includes robots? (0.86,0.28,0.55)

  0 shows/movies

  1-5 shows/movies

  6-10 shows/movies

  10+ shows/movies

- Have you ever interacted with a robot? (select all that apply) (0.22,0.22,0.22)

  Museum or theme park animatronics

  Toys such as Furby

  Robot vacuum

  Classroom robots or Battlebots

  Sawyer (the robot in this experiment)

  Everyday items such as cell phone, computer, ATM, or Xbox

  Other

- Have you ever built a robot? (select all that apply) (0.11,0.22,0.11)

  Classroom setting

Club setting

Other

- Have you ever controlled a robot? (select all that apply) (0.33,0.11,0.22)

  Teleoperation or remote control

  Speech, Gesture, Commands

  Computer programmed

  Other

A.2.1.2 Attitudes Towards Robots

The next set of questions detailed participants' attitudes towards robots in general. All questions were on a 7-point Likert scale, with 1 being Strongly Disagree and 7 being Strongly Agree. p-values in this section are based on the difference between pre- and post-activity surveys.

- I would feel uneasy if robots really had emotions. (0.27,0.14,0.92)
- Something bad might happen if robots developed into living beings. (0.12,0.95,0.21)
- I would feel relaxed talking with robots. (0.86,0.76,0.98)
- I would feel uneasy if I was given a job where I had to use robots. (0.003,0.06,0.45)
- If robots had emotions I would be able to make friends with them. (0.88,0.71,0.95)
- I would feel nervous operating a robot in front of other people. (0.02,0.84,0.06)
- I would hate the idea that robots were making judgements about things. (0.58,0.58,1.0)
- I would feel very nervous just standing in front of a robot. (0.26,1.0,0.26)
- I feel that if I depend on robots too much, something bad might happen. (0.71,0.99,0.78)
- I am good at working with robots. (0.39,1.0,0.39)

- I would feel paranoid talking with a robot. (0.98,0.58,0.68)
- I am concerned that robots would be a bad influence on children. (0.21,0.34,0.95)
- I feel that in the future society will be dominated by robots. (0.58,0.94,0.78)
- Most robots make poor teammates. (1.0,0.96,0.96)
- Most robots possess adequate decision making capabilities. (0.16,0.37,0.85)
- Most robots are easy to understand. (0.8,0.34,0.7)

A.2.1.3 Attitudes Towards Sawyer

This section of questions pertained to the participants' initial impression of the Sawyer robot. All questions are on a 7-point Likert scale. 1 was the adjective on the left, 7 was the adjective on the right. p-values in this section are based on the difference between pre- and post-activity surveys.

- I [blank] Sawyer. (Like/Dislike) (0.89, 0.97, 0.97)
- Sawyer is: (Unkind/Kind) (0.006, 1.0, 0.44)
- Sawyer is: (Ignorant/Knowledgeable) (0.07, 1.0, 0.07)
- Sawyer is: (Incompetent/Competent) (0.29, 0.92, 0.15)
- Sawyer is: (Unintelligent/Intelligent) (0.59, 0.98, 0.47)
- Sawyer is: (Foolish/Sensible) (0.31, 0.67, 0.07)
- Sawyer is a(n): (Individualist/Team Player) (0.66, 0.03, 0.15)
- Sawyer is: (Unlikeable/Likeable) (0.1, 0.9, 0.2)
- Sawyer is: (Unfriendly/Friendly) (0.53, 0.7, 0.16)
- Sawyer is: (Stubborn/Agreeable) (0.04, 0.52, 0.29)

A.2.2 Inter-Round Survey Questions

Other than the first question, which asked participants to select the round they had just completed, questions were on a 7-point Likert scale, and values for 1 and 7 are indicated in the form (adjective for 1 / adjective for 7) p-values in this section are written in the form (optimal r1/r2, optimal r1/r3, optimal r2/r3, median r1/r2, median r1/r3, median r2/r3, PACT r1/r2, PACT r1/r3, PACT r2/r3)

- Round

  1

  2

  3

- How mentally demanding was the task? (Very Low Mental Demand/Very High Mental Demand) (0.9, 0.9, 0.9, 0.83, 0.9, 0.9, 0.9, 0.9, 0.9)

- How successful were you in accomplishing what you were asked to do? (Perfect / Complete Failure) (0.9, 0.9, 0.9, 0.9, 0.9, 0.9, 0.9, 0.9, 0.9)

- How hard did you have to work to accomplish your level of performance? (Very Low Effort / Very High Effort) (0.72, 0.9, 0.8, 0.83, 0.9, 0.9, 0.75, 0.9, 0.9)

- How discouraged, irritated, stressed, and annoyed were you? (Very Low Frustration / Very High Frustration) (0.67, 0.53, 0.9, 0.82, 0.82, 0.9, 0.84, 0.9, 0.9)

- I was confident that Sawyer would choose the same block that I chose. (Very Low Confidence / Very High Confidence) (0.78, 0.56, 0.23, 0.9, 0.09, 0.17, 0.75, 0.16, 0.48)

- I understand how Sawyer was choosing blocks. (No Understanding / Complete Understanding) (0.79, 0.79, 0.44, 0.85, 0.65, 0.36, 0.82, 0.42, 0.75)

A.2.3 Post-Activity Survey

Listed p-values are of the form (conventions/median, conventions/optimal, median/optimal).

A.2.3.1 Game Comprehension

These questions concerned participants' understanding of the game. All questions are on a 7-point Likert scale. Value labels were Strongly Disagree (1) and Strongly Agree (7) unless otherwise stated.

- I understood the rules of the game. (0.9, 0.9, 0.9)
- I used the previous selections shown on the tablet to make my decisions. (0.28, 0.9, 0.44)
- I knew things about the game that Sawyer didn't know. (0.9, 0.37, 0.32)
- I understood the goal of the game. (0.9, 0.81, 0.86)
- I kept track of our score at each turn. (0.9, 0.9, 0.9)
- Sawyer knew things about the game that I didn't know. (0.79, 0.9, 0.79)
- How much did your team's score influence the decisions you made? (No Influence / Score Was the Only Influence) (0.66, 0.54, 0.9)

A.2.3.2 Attitudes Towards Sawyer

The questions in this section were identical to those asked in the same section in the Pre-Activity Survey.

A.2.3.3 Team Fluency and Performance

These questions concerned participants' perceptions of their team. All questions are on a 7-point Likert scale. Value labels were Strongly Disagree (1) and Strongly Agree (7) unless otherwise stated.

- The robot and I contributed equally to the success of the team. (0.9, 0.6, 0.74)
- Working with Sawyer was stressful or frustrating. (0.31, 0.67, 0.75)
- I am responsible for the team's score. (0.9, 0.9, 0.9)
- The team worked fluently together. (0.24, 0.07, 0.82)
- I helped the robot accomplish the task. (0.9, 0.24, 0.46)
- The team's coordination improved over time. (0.9, 0.02, 0.04)
- The robot was cooperative. (0.26, 0.47, 0.87)
- The robot is responsible for the team's score. (0.75, 0.41, 0.14)
- If I were a robot, the team would have scored better. (0.9, 0.59, 0.61)
- The robot perceived accurately what I was trying to do. (0.9, 0.72, 0.86)
- I am good at working with robots. (0.53, 0.82, 0.24)
- I contributed more to the success of the team. (0.83, 0.26, 0.59)
- Working with Sawyer was difficult. (0.74, 0.25, 0.66)
- The robot and I were working toward the same goal. (0.9, 0.9, 0.9)
- The robot helped me accomplish the task. (0.9, 0.36, 0.58)
- Sawyer is good at working with humans. (0.41, 0.56, 0.9)
- I find what I am doing with the robot confusing. (0.9, 0.9, 0.9)
- I was a good teammate to Sawyer. (0.075, 0.041, 0.9)
- There was a team leader (True/False multiple choice) (0.77, 0.9, 0.9)

- If there was a team leader, who was the team leader? (If there was no team leader, skip this question) (Sawyer/Me) (0.56, 0.56, 0.56)
- The robot contributed more to the success of the team. (0.86, 0.56, 0.29)
- Over time, the way I selected blocks changed. (0.9, 0.37, 0.35)
- Who is more responsible for the team's success or failure? (Sawyer / Me) (0.79, 0.9, 0.82)
- Sawyer was a good teammate to me. (0.11, 0.04, 0.9)
- I would have scored better if my teammate was human. (0.018, 0.004, 0.9)
- I would work with Sawyer again. (0.35, 0.06, 0.66)

A.2.3.4 Robot Predictability and Understandability

The questions in this section relate to the participant's understanding of the robot and how predictable they found the robot. All questions were on a 7-point Likert scale from Strongly Disagree to Strongly Agree unless othwerwise indicated.

- Sawyer was unpredictable. (0.9, 0.014, 0.0395)
- I understood why Sawyer made the decisions it did. (0.63, 0.0235, 0.18)
- The way Sawyer selected blocks was unclear to me. (0.35, 0.0078, 0.2)
- I could easily predict what block Sawyer would pick next. (0.31, 0.0078, 0.24)
- The way Sawyer picked blocks made sense to me. (0.39, 0.0069, 0.16)
- As the game progressed, I was more easily able to predict which block Sawyer would pick next. (0.9, 0.001, 0.001)
- Sawyer's decisions didn't make sense. (0.64, 0.07, 0.39)
- Sawyer picked the best block for the team. (0.07, 0.16, 0.85)

- Sawyer chose blocks randomly. (0.24, 0.001, 0.008)
- Most people would be able to understand how Sawyer made decisions. (0.17, 0.01, 0.47)
- I chose blocks (intuitively / analytically) (0.63, 0.24, 0.045)
- Fill in the blank: By the end of Round [blank] I could easily predict which block Sawyer would pick next. (multiple choice)

  1

  2

  3

  None

A.2.4 Round 4 Survey

For this survey, participants were shown 5 novel game boards and were asked the same set of multiple choice questions for each of them. Participants were instructed not to guess, and to select "unsure" if they were not totally certain about their answer.

- Which color is the block Sawyer will pick first?

  blue

  red

  yellow

  unsure

- Which shape is the block Sawyer will pick first?

  circle

  triangle

  square

unsure

- Which color is the block Sawyer will pick last?

  blue

  red

  yellow

  unsure

- Which shape is the block Sawyer will pick last?

  circle

  triangle

  square

  unsure

black

Appendix B

PRESTO Appendix

B.1 PRESTO Appendix

Questions given between rounds about predictability are as follows:

- I was confident that the robot would go where I thought it would go.
- I understand how the robot was making decisions.

Post-survey questions about predictability and understandability are as follows:

- The robot was unpredictable.
- I understood why the robot made the decisions it did.
- The way the robot made decisions was unclear to me.
- I could easily predict where the robot would go next.
- The way the robot moved made sense to me.
- As the game progressed, I was more easily able to predict where the robot would go next.
- The robot's decisions didn't make sense.
- The robot picked the best path for the team.
- The robot moved randomly.

- Most people would be able to understand how the robot made decisions.
- I moved (intuitively/analytically).
- Fill in the blank: By the end of Round [BLANK] I could easily predict where the robot would go next.

Modified and additional team fluency questions are as follows:

- If I were a robot, the team would have scored better.
- Over time, the way I made decisions changed.
- I would have scored better if my teammate was human.