December 2019

# Human-Machine Teamwork: An Exploration of Multi-Agent Systems, Team Cognition, and Collective Intelligence

Lorenzo Barberis Canonico
*Clemson University*, lorenzb@g.clemson.edu

# Human-Machine Teamwork: An Exploration of Multi-Agent Systems, Team Cognition, and Collective Intelligence

A Dissertation
Presented to
the Graduate School of
Clemson University

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy
Human-Centered Computing

by
Lorenzo Barberis Canonico
December 2019

Accepted by:
Dr. Nathan McNeese, Committee Chair
Dr. Brian Dean
Dr. Kelly Caine
Dr. Guo Freeman
Dr. Richard Pak

# Abstract

One of the major ways through which humans overcome complex challenges is teamwork. When humans share knowledge and information, and cooperate and coordinate towards shared goals, they overcome their individual limitations and achieve better solutions to difficult problems. The rise of artificial intelligence provides a unique opportunity to study teamwork between humans and machines, and potentially discover insights about cognition and collaboration that can set the foundation for a world where humans work with, as opposed to against, artificial intelligence to solve problems that neither human or artificial intelligence can solve on its own.

To better understand human-machine teamwork, it's important to understand human-human teamwork (humans working together) and multi-agent systems (how artificial intelligence interacts as an agent that's part of a group) to identify the characteristics that make humans and machines good teammates. This perspective lets us approach human-machine teamwork from the perspective of the human as well as the perspective of the machine. Thus, to reach a more accurate understanding of how humans and machines can work together, we examine human-machine teamwork through a series of studies.

In this dissertation, we conducted 4 studies and developed 2 theoretical models:

First, we focused on human-machine cooperation. We paired human participants with reinforcement learning agents to play two game theory scenarios where individual interests and collective interests are in conflict to easily detect cooperation. We show that different reinforcement models exhibit different levels of cooperation, and that humans are more likely to cooperate if they believe they are playing with another human as opposed to a machine.

Second, we focused on human-machine coordination. We once again paired humans with machines to create a human-machine team to make them play a game theory scenario that emphasizes convergence towards a mutually beneficial outcome. We also analyzed survey responses from the participants to highlight how many of the principles of human-human teamwork can still occur in human-machine teams even though communication is not possible.

Third, we reviewed the collective intelligence literature and the prediction markets literature to develop a model for a prediction market that enables humans and machines to work together to improve predictions. The model supports artificial intelligence operating as a peer in the prediction market as well as a complementary aggregator.

Fourth, we reviewed the team cognition and collective intelligence literature to develop a model for teamwork that integrates team cognition, collective intelligence, and artificial intelligence. The model provides a new foundation to think about teamwork beyond the forecasting domain.

Next, we used a simulation of emergency response management to test the different teamwork aspects of a variety of human-machine teams compared to human-human and machine-machine teams. Lastly, we ran another study that used a prediction market to examine the impact that having AI operate as a participant rather than an aggregator has on the predictive capacity of the prediction market.

Our research will help identify which principles of human teamwork are applicable to human-machine teamwork, the role artificial intelligence can play in enhanc-

ing collective intelligence, and the effectiveness of human-machine teamwork compared to single artificial intelligence. In the process, we expect to produce a substantial amount of empirical results that can lay the groundwork for future research of human-machine teamwork.

# Dedication

This dissertation is dedicated to the people in my life whose support has made and continues to make the American Dream a future I get to passionately chase every day.

Andres, I dedicate this dissertation to you because nothing I will ever do, academic or otherwise, will ever truly be just my own: it will always be something we have accomplished together – as a team! From the moment we've met, our friendship has redefined every aspect of our lives by unlocking potential that makes anything feel possible as long we do it together. We are on a mission that is going to demand everything of us, and no one but you could make me feel like we've got this because you are and always will be the Hermano.

Mom and Dad, I dedicate this dissertation to you as well. I will never settle until I find a way to honor the tremendous sacrifices you have both made to give me the opportunity to pursue a life in the most exceptional country on the planet. The PhD is just another entry in the long-running list of unique opportunities I will seize for the rest of my life in order to thank you for betting it all on your wild child.

# Acknowledgments

Clemson University and the School of Computing, for welcoming me at a time when all doors seemed locked shut. The Human-Centered Computing program has been one of the most creatively empowering experiences of my life, much of which is directly attributable to the interdisciplinary culture that emerges from such interesting people working on helping humans thrive through technology.

My dissertation committee (Brian Dean, Kelly Caine, Guo Freeman and Richard Pak), for committing with me to the wild ride that produced this dissertation. I don't take that for granted, so I am extremely grateful for your time and support. Specifically, thank you for challenging me to consider all the angles: Dr. Pak's advice on study design set the NeoCITIES studies in motion. Dr. Freeman's insight into the ubiquity of collective intelligence outside of traditional teamwork areas forced me to tackle a dissertation broader in scope. Dr. Caine's research philosophy will be something I will carry with me for the rest of my life: you have renewed my faith in science despite the replication crisis and the distorted incentive structures of academia. Finally, Dr. Dean for pushing me to delve deeply into the technical side: whether through our discussions about game theory and matching algorithms or cellular automata and tangles, you have shown me that computer science holds many of the secrets to attaining the technological leverage necessary to tackle some of humanity's toughest challenges.

Dr. Safro and Dr. Knijnenburg, whose courses have trained me in the complex methodologies behind this dissertation. You were both right: I learned far more than I ever imagined despite my initial struggles.

The TRACE Lab, for creating a culture of continuous improvement that is enabling us to always discover new ways to be more productive. Raf Dejesus, for helping me grow by refusing to settle for anything less than the best version of myself as a leader. Bekk Blando, for teaching me the value of patience, especially when caring about others. Rui Zhang, for showing me how asking the right questions is often just as important as finding the answers. Anurata Hridi, for indulging me far too often in our crazy conversations about game theory and swarm intelligence. Jake Armstrong, for helping me confront the reality that I too easily give up on excellence and attention to detail when pressed for time. Mark Blasko, for always nudging me to exercise: you've taught me that exercise is not just a past time, but can be a repeated behavior with which I can shape who I become. Steve Russell, for making me believe in the synergy between graduate students and undergraduate students: you show how real talent engages in multiple domains.

Even though they are also part of the TRACE Lab, Beau Schelble and Chris Flathmann deserve their own special acknowledgement. Whether late night or the day before submitting my dissertation, you both helped follow through on an ambitious goal. Working together on all of our projects has been one of the most energizing parts of my entire time as a PhD student. Our teamwork is the best kind of teamwork: the kind that makes over-promising AND over-delivering possible.

Last but not least, I owe a huge debt of gratitude to the man who made this entire PhD possible: Nathan McNeese. He was a phenomenal advisor, is a first-class teammate, and will be a source of creative inspiration for my entire life. He made me a believer in the underestimated power of human cognition and in the importance of maturity for leadership. Thank you for taking a chance on me as your first PhD student: hopefully working together changed your life as much as it did mine.

# Table of Contents

# List of Tables

# List of Figures

# Chapter 1

# Introduction

## 1.1 Overview

The goal of this dissertation is to advance our understanding of human-machine teamwork by studying coordination and cooperation between humans and artificial intelligence (AI). We examine human-machine teamwork from the perspective of both the human and the machine by exploring the connections between team cognition and collective intelligence as they apply to humans and AIs interacting in various ways (AI as a coordinator, AI as a team-mate, etc) and various scales (small teams, large complex systems, dyadic partnerships). We first review the literature on human-human teamwork, reinforcement learning, and collective intelligence in order to identify the key concepts needed to ground human-machine teamwork. Secondly, we implement two different experiments using different models of human-machine teamwork inspired by these principles. Lastly, we use our results to make inferences about the ways in which human-machine teamwork manifests in multiple scales (small teams vs large crowds) and dimensions (AI playing different roles as part of the team).

## 1.2 Problem Motivation

One of the major forces that will influence the course of humanity in the next century is the development of artificial intelligence (AI). Specifically, the primary objective of contemporary artificial intelligence research is the development of artificial general intelligence (AGI): a strong AI that would theoretically be capable of matching and eventually outperforming human intelligence across a variety of different tasks where intelligence plays a decisive role [Goertzel and Pennachin, 2007]. AGI differs from narrow AI in that it holds a true understanding of itself and its environment that makes it capable of easily transferring learning between different tasks [Pennachin and Goertzel, 2007]. A major unfortunate consequence of this pursuit is the attempt at rendering human effort obsolete by automating away the roles humans play in complex processes. However, more closely examining human-machine partnerships at both the micro and macro level suggests that prioritizing joint efforts between humans and AIs will result in much higher degrees of effectiveness and progress.

At the micro level, prior research suggests that effective human-machine partnerships outperform not just human teams, but also sophisticated AI systems. One of the major areas where this insight is self-evident is the healthcare domain. For example, in cancer detection, teams of doctors partnering with machine learning algorithms outperformed both expert teams as well as state-of-the-art neural networks in diagnosing cancer [Wang et al., 2016]. Specifically, although the AI's error rate was 2.9% when compared to the average pathologist's 3.5%, the pair working together had a collective error rate of 0.5%. In this context, five times as many effective diagnoses were performed by the human-machine team than by the AI alone, potentially impacting the lives of the many people afflicted with cancer each year. These results are not outliers, and are part of a broader pattern. For instance, recent studies of children's infections, which are notoriously challenging to diagnose, showed that an AI processing the patient's electronic health records had an 86.2% accuracy rate when compared to the average experienced doctor's 75% rate. However, tuning the algorithm to minimize false negatives at the expense of minimizing false positives, and having the doctors integrate their own perception and judgment in the diagnosis led to a 99.4% detection rate of deadly infections that claim the lives of children every year. In both instances, the AI's performance alone would have justified the incorrect inference that the doctor was no longer necessary, which would have resulted in massive losses in accurate diagnoses by failing to take into account the possibility of humans and AIs working together.

At the macro-level, the interactions between large groups of humans and AIs give rise to complex systems that can be challenging to manage. For example, in stock markets and financial marketplaces, “Flash Crashes” can have disastrous consequences. A ”Flash Crash” occurs when financial security prices collapse rapidly within a short-time window [Bozdog et al., 2011]. On May 6, 2010, all major stock market indexes collapsed within a 36 minute window, with the Dow Jones Industrial Average having its biggest intraday decline in history [Kirilenko et al., 2017]. Prior to the unprecedented rebound, over $1 trillion in market capitalization was lost [Grocer, 2010]. Similarly, on October 7, 2016, the value of the British pound fell over 6%, which put it at its lowest level against the dollar since May 1985 [Ismail and Mnyanda, 2016]. These extreme events were the result of High Frequency Trading (HFT) algorithms reacting to each other's buy and sell orders, which in turn triggered a negative feedback loop that exaggerated a downward market move. These kinds of chaotic chain reactions can lead to catastrophic consequences for any complex system that does not take into account the feedback loops between human and machine behavior. Furthermore, whether humans and AIs can cooperate at a larger scale is still an open question. Recent research published in Nature shows how bots attempting to solve a graph coloring problem, a matching task that requires high levels of coordination, fail to achieve a globally optimal solution [Shirado and Christakis, 2017]. Essentially, the bots can successfully coordinate on a micro-level but fail to do so on a macro-level, thereby preventing the collective goal of a stable match to occur. However, the study also points to a way to transcend such limitations: adding humans to the team and nudging attempts at coordination through a separate type of bot that pushes the humans and AIs to communicate even at large scales [Shirado and Christakis, 2017].

All of these examples hold a key characteristic in common: the initial results in AI performance would have justified the erroneous conclusion that human expertise should be disregarded because of superior AI performance, when in fact ensembling the two results in major improvements compared to AI performance alone. These successes in having humans and AIs work together, at both the micro and macro level, can be attributed to the unique advantages that emerge from harvesting human and AI potential in a compatible and integrated way, and should pave the way for more research of this kind. In essence, studying human-machine teamwork enables Human Computer Interaction (HCI) to shape the future of AI development in a human-centered direction.

## 1.3 Research Motivations

AI has the ability to contribute in many different facets of teamwork by playing a variety of roles such as coordinator, decision-aid, agent, and teammate. Therefore, human-machine teamwork encompasses a wide variety of sub-disciplines that sit at the intersection between cognitive science and computer science. Thus human-machine teamwork in all of its forms (micro vs macro, autonomous vs automated) can only be understood through an interdisciplinary perspective. To that end, we identify the two core research fields that are needed to understand the role AI can play in teamwork, as well as the theoretical and empirical gaps that exist when it comes to human-machine teamwork specifically.

### 1.3.1 Human-Human Teamwork

Human teamwork is the obvious starting point to ground research in human-machine teamwork. Decades worth of prior research have begun to understand the principles that govern human teamwork, with concepts such as shared situational awareness, transactive memory, and team cognition [Cooke et al., 2012]. However, human-machine teamwork has been comparatively under-researched because of strong technical limitations that have historically been encountered when building agents [?]. Up until recently, most research was restricted to either the robotic domain or contexts where the human had a supervisory role on an automated system [Gao et al., 2016, Sheridan and Telerobotics, 1992]. Despite those limitations however, breakthroughs in cognitive modeling and artificial intelligence recently opened up the opportunity to study humans interacting with machines at a peer-level [Sun et al., 2006].

Yet, only a limited amount of empirical research has been conducted on humans interacting with autonomous (as opposed to automated) agents in a team setting, especially looking at the presence of characteristics of successful human teamwork [Demir et al., 2017]. Specifically, whether the assumptions of the human-human teamwork paradigm will end up being validated in human-machine teams is now an open question, as contemporary research has begun to question whether human teamwork should be the model that drives the design and experimentation of human-machine teams in the first place [McNeese et al., 2019]. Thus, a major gap exists in the literature with regards to empirically testing whether the principles of human-human teamwork will prove as decisive in human-machine team performance as they are in human team performance.

### 1.3.2 Collective Intelligence

Collective intelligence (also known as "the wisdom of the crowd") refers to the emergence of a globally optimal solution from the proper aggregation of individual information that exceeds what any individual would be capable of on their own [Watkins, 2007]. Collective intelligence can emerge in large crowds of humans correctly guessing the number of jelly beans in a jar, in complex multi-agent systems of AIs solving complex resource allocation problems, and in humans and AIs working together to solve graph coloring problems [Surowiecki, 2005, Shirado and Christakis, 2017]. Thus, collective intelligence can be understood as the large scale equivalent of many of the emergent cognitive properties of good teamwork.

The major research avenue for collective intelligence is prediction markets: mechanisms that enable participants to bet on future events in the way stock traders place bets in the stock market in anticipation of corporate earnings [Surowiecki, 2005]. The way collective intelligence manifests itself in prediction markets is through remarkable forecasting accuracy, far above that of individual or expert level judgement. Studies of the use of prediction markets in sports [Goel et al., 2010] and politics [Hanson, 2003] have shown orders of magnitude improvements in forecasting accuracy between other methodologies and prediction markets.

However, prediction markets are far from perfect. Even though they exhibit high degrees of signal processing by compensating for overconfidence as well as underconfidence, they still face severe limitations [Hanson, 2003]. Specifically, major biases such as sampling error, market maker bias, and convergence error have been identified as unique barriers to collective intelligence emerging [Dudik et al., 2017]. The majority of prior research has focused on addressing these issues through the design of better calibration methods, which unfortunately have turned out to be unsuccessful [Chen et al., 2005].

Furthermore, comparatively little research exists on the role AI can play in improving the collective intelligence of a prediction market. The few studies that exist on the subject focus on developing randomized bots whose trading patterns induce human participants to improve their forecasts [Malone, 2018]. This gap presents a major research opportunity to integrate the recent advancements in machine learning into prediction markets. Specifically, there are two major ways in which AI can be studied: hybrid prediction markets (where AIs and humans participate at a peer-level) and machine-learning-based aggregators (AI receiving a prediction market's trading data as its input in order to produce more accurate predictions).

Thus, a major gap exists in the literature with regards to how humans and AI can interact as a team at a large scale in order to produce collective intelligence.

### 1.3.3 Research Objective & Questions

The main objectives of this dissertation are to (a) identify the similarities and differences between human-machine teamwork, human-human teamwork, and machine-machine teamwork; (b) explore the unique ways in which team cognition emerges in human-machine teams; (c) identify the ways in which AI can enhance collective intelligence in human-machine teamwork; and (d) develop empirically-backed design guidelines for the integration of AI in prediction markets;

These following main research areas and research questions will address the gaps in the research and this dissertation's main objectives:

*Understanding team cognition in human-machine teams*

RQ1: Which principles of human teamwork are applicable to human-machine teams?

RQ2: In what ways do human-machine teams outperform machine-machine teams?

RQ3: What are the performance tradeoffs between human-human, human-machine, and machine-machine teams?

*Understanding collective intelligence in human-machine teamwork*

RQ4: Do human-machine multi-agent systems exhibit higher degrees of collective intelligence than human-only systems?

RQ5: Can collective intelligence be used as the input to artificial intelligence?

RQ6: Which machine learning approaches are best suited for prediction markets?

| **Research Gaps** | **Research Questions** |
|---|---|
| There is a lack of understanding with regards to the extent to which team cognition is possible in human-machine teams | RQ1 and RQ2 |
| There is a lack of clarity as to what principles of human-teamwork apply to human-machine teams | RQ1 |
| Few guidelines exist as to how to optimally configure human-machine teams | RQ2 and RQ3 |
| The literature has tended to restrict AI research to the purpose of automating human effort | RQ2 and RQ5 |
| Very little research exists studying the use of artificial intelligence to enhance collective intelligence of both human crowds as well as human-machine teams | RQ4, RQ5 and RQ6 |

Table 1.1: Research Gaps and Research Questions

To better understand how these research questions were developed based on the gaps within the literature; the following tables have been constructed 1.1 which links the research questions to the specific research gap.

### 1.3.4 Overview & Summary of Studies

To achieve the research objectives outlined previously, we conducted 4 studies and developed 2 new models from the literature:

**Study 1: Game Theory for Teams**. Recent advancements in reinforcement learning have expanded the type of models available to conduct agent-based machine learning research. To that end, many of these models have not been studied from an HCI perspective. This study employs a new methodology that leverages game theory from an HCI perspective to investigate the extent to which both AI and human players alter their strategic behavior when interacting in cooperative and non-cooperative scenarios. After conducting a large scale study of human and AI gameplay across multiple game theory models, the results indicate that different reinforcement learning models cooperate differently with humans, and that human players display biases towards AI that influence the extent to which they end up cooperating in social dilemmas [Barberis Canonico et al., 2019d].

**Study 2: Human-Machine Teams as Multi-Agent Systems**. Understanding human-machine teams requires bridging the gap between human teamwork and multi-agent systems. To that end, we conducted a study where human participants were paired up with reinforcement learning agents to play a game theory scenario that emphasizes cooperation. Results indicate that human biases in favor or against AI have significant effects on aggression and peacefulness, and that shared understanding plays a major role in cooperation [Barberis Canonico et al., 2019b].

Both of these studies served as the foundation for our approach to human-machine teamwork by leveraging game theory models to observe how machines change their cooperative behavior when playing with humans and vice versa. Our results informed our understanding that many team cognition aspects can emerge from human-machine teams even when communication is not possible.

**Model A: Human-Centered Prediction Markets**. There is an ever-growing literature on the power of prediction markets to harness "the wisdom of the crowd" from large groups of people; however, traditional prediction markets are not designed in a human-centered way, often restricting their own potential. This creates the opportunity to implement a cognitive science perspective on how to enhance the collective intelligence of the participants. Thus, we proposed a new model for prediction markets that integrates human factors, cognitive science, game theory, and machine learning to maximize collective intelligence. We do this by identifying the connections between prediction markets and collective intelligence, using human factors techniques to analyze our design, and outlining the practical ways with which our design enables artificial intelligence to complement human intelligence [Barberis Canonico et al., 2019c].

**Model B: Collectively Intelligent Teamwork**. We proposed a new model for teamwork that integrates team cognition, collective intelligence, and artificial intelligence. We did this by first characterizing what sets team cognition and collective intelligence apart, and then reviewing the literature on "superforecasting" and the ability for effectively coordinated teams to outperform predictions by large groups. Lastly, we delved into the ways in which teamwork can be enhanced by artificial intelligence through our model, highlighting the many areas of research worth exploring through interdisciplinary efforts [Barberis Canonico et al., 2019a].

Both of the models are currently published conceptual models and this dissertation empirically tested them through the 2 studies outlined below:

**Study 3: Team Cognition in Human-Machine Teams**. We tested the team cognition aspects (shared understanding, shared situational awareness, etc) of a variety of human-machine teams through a simulation of emergency response management. We also compared human-machine teams' performance against that of human-only and machine-only teams, while also gaining a qualitative understanding of the factors examined in Study 1 and 2.

**Study 4: Human-Machine Collective Intelligence in Prediction Markets**. We implemented model A to empirically test the model and gain insight into the role artificial intelligence can play in the enhancement of collective intelligence. We studied prediction markets as a human-machine teamwork problem by looking at the impact that having AI operate as a participant rather than an aggregator has on the predictive capacity of the prediction market.

Overall, we sought to study the cognitive implications of human-machine teamwork at the micro level through team cognition and at the macro level through collective intelligence. The insight gained through these studies will apply to human teamwork and multi-agent systems by covering both perspectives through the conceptual basis of human teamwork and the computational foundation of contemporary reinforcement learning.

### 1.3.5 Conclusion

Research on human-machine teamwork is critical to alter the trajectory of AI research towards enhancing human potential by leveraging the best features of each cognition: human cognition as well as machine cognition. Given that human-machine interactions can occur at multiple levels of decision-making, as well as in a variety of goal-oriented settings, it's important to establish strong methodological foundations that can yield insights into how to develop AI in a way that optimizes the human-machine partnership. In the process, it's very likely that valuable contributions will be made to multiple areas of science, ranging from HCI to cognitive science.

# Chapter 2

# Background and Related Work

## 2.1 Overview

Pursuant of the interdisciplinary nature of this proposal, a wide variety of theoretical constructs need to be established to ground our perspective on human-machine teamwork. This chapter establishes the theoretical foundations of human-machine teaming (the current state of the research), reinforcement learning (the machine learning paradigm underlying AI in all of our studies), team cognition (the principal concept in Study 3), and collective intelligence (the basis for Study 4). Each theory explains a particular emergent phenomenon from different ensembles of agents (human and AI) as they engage in a shared form of information processing that cannot be understood at the individual level. For each theory, we begin by summarizing the literature as it applies to groups of humans, and then identify models and concepts from the theory that can extend and generalize to human-machine teams.

## 2.2 Human-Machine Teamwork

### 2.2.1 HAT: Human-Autonomy Teaming vs Human-Automation Teaming

Traditional perspectives on human-machine teamwork have evolved out of cognitive science and human factors research in the interaction design and evaluation of humans interacting with different degrees of automated technology [Vagia et al., 2016, Demir et al., 2017]. A new perspective was introduced when technology reached a point where it could function independently of human action, thus leading to an automation-vs-autonomy divide. The divide is understood to be predicated upon the degree to which human control determines the technology's actions: automation refers to systems that exactly follow their programmed instructions without independent action, whereas autonomy refers to systems who make their own task-related decisions independently of human control [Vagia et al., 2016]. A more contemporary definition of autonomy is given by Endsley (2015), as a system with intelligence-based capabilities to respond to situations outside of the range of the possibilities considered at the design phase. Much of the prior work in this area has been related to human-robot teaming and human control of automation in a supervisory role [Gao et al., 2016, Sheridan and Telerobotics, 1992]. However, new technologies have expanded the realm of possible machine teammates humans can interact with (Ex. hazard warning systems, virtual agents, decision-support systems, etc) [McNeese et al., 2019].

From a human teamwork perspective, human-autonomy teaming is most analogous to human-human teamwork because both AI and human teammates interact through a peer-relationship as opposed to a master-subordinate relationship (Adler, 1997). Therefore, human-machine teaming originates from the human-autonomy literature because both conceptual models share the basic characteristic of humans and machines engaging in interdependent behaviors towards a common goal [McNeese et al., 2019].

### 2.2.2 Major Results in Human-Machine Teamwork

Prior research has approached the question of human-machine teamwork through several different methodologies. For example, Fan et al (2010) developed an agent based on the naturalistic decision-making model of cognition, and their findings suggest that the agent can help human decision-makers by improving situational awareness albeit at the expense of higher cognitive load. Identifying these trade-offs is an important step to develop a precise understanding of the ways in which human-machine teams differ from human-human teams. Other research by Chiou and Lee (2016) have demonstrated that the cooperative behavior of machine agents in joint resource management and scheduling tasks has a meaningful impact on overall team performance. More recently, comparative studies of human-machine teams and human-human teams in a simulation highlighted team-level communication and coordination deficiencies in human-autonomy teams [McNeese et al., 2019]. The same methodology has produced findings highlighting the importance of situational awareness and team synchrony in human-autonomy teaming as well [McNeese et al., 2018, Ball et al., 2010].

At a higher level, Klein et al (2004) developed a series of principles to set up robust foundations for human-machine teams:

1. *To be a team player, an intelligent agent must fulfill the requirements of a Basic Compact to engage in common-grounding activities*

2. *To be an effective team player, intelligent agents must be able to adequately*

*model the other participants' intentions and actions vis-à-vis the joint activity's state and evolution*

3. *Human-agent team members must be mutually predictable.*

4. *Agents must be directable*

5. *Agents must be able to make pertinent aspects of their status and intentions obvious to their teammates*

6. *Agents must be able to observe and interpret pertinent signals of status and intentions*

7. *Agents must be able to engage in goal negotiation.*

8. *Support technologies for planning and autonomy must enable a collaborative approach*

9. *Agents must be able to participate in managing attention*

10. *All team members must help control the costs of coordinated activity*

They summarized much of the literature on human-machine teamwork not just by identifying these principles, but also delving deeply into why they are so important to the effectiveness of human-machine teams. They point out that team coordination between humans and machines can only occur if there is a commitment to goal alignment that can facilitate coordination towards shared goals [Klien et al., 2004]. Predictability also plays a major role because many team-related actions rely on a highly interdependent set of activities that are only possible if the team has developed shared knowledge through extended experience in working together [Klien et al., 2004]. Directability on the other hand refers to the ability to assess and change actions to adapt to a dynamic environment, and it plays a key role in the team's ability to effectively respond to activities that require high levels of coordination [Christoffersen and Woods, 2002]. This principle has led researchers to focus on developing ways for humans to control aspects of agent autonomy in a way that can be dynamically specified and easily understood [Christoffersen and Woods, 2002].

Klein et al's (2004) model for human-machine teamwork has profoundly shaped many of the design choices of agents for human-machine teams. Specifically, prior research in mixed human-robotics team used these principles to design a general-purpose agent to manage collaboration and support human-agent, human-human, and agent-agent interactions [Allen and Ferguson, 2002]. Yet, despite these results, progress in the area of human-machine teamwork has been limited by several factors.

### 2.2.3 Limitations and Research Gaps

The primary barrier to human-autonomy research has been technical. The development of highly autonomous agents that can effectively interact with humans and adapt to dynamic environments has proven non-trivial across a variety of domains [Klien et al., 2004]. It's been particularly challenging to study team-related concepts that involve high degrees of communication because natural language processing, especially at the sophisticated level required to match human-human communication, is still an open problem [Bates and Weischedel, 2006]. However, recent breakthroughs in reinforcement learning (highlighted in the next section) provide a new opportunity to build sophisticated, general-purpose AIs.

On the other hand, many of the research gaps highlighted in the Introduction still exist because our understanding of human-human teamwork grows more sophisticated with the passage of time, generating a wealth of concepts, models, and principles that may or may not apply to human-machine teams and have yet to be studied [McNeese et al., 2019]. Fundamentally, it is still not clear what many of the similarities and differences between human-human teams and human-autonomy teams are, and only a small amount of evidence exists actually validating the assumption that human teamwork should be the basis for our understanding and design of human-machine teams [McNeese et al., 2019]. For example, even though team situational awareness and interactive team cognition have been identified as critical components of successful human-human teams, very little empirical evidence suggests this is the case for human-machine teams [Demir et al., 2017]. Prior research in command-and-control environments suggests that it's very likely that human-machine teamwork requires the human to compensate for the machine's shortcomings and unpredictability through specific mental models [Endsley, 2015].

### 2.2.4 Relationship to the Dissertation

This dissertation addresses many of the historical limitations of HAT research by incorporating reinforcement learning (explained in the next section) as the machine learning paradigm to build teammates for human-machine teams. This decision will establish a methodology that can generalize to a variety of settings, unlike prior research.

Furthermore, the invitation by the current human-machine literature to explore alternatives to the human-human model of teamwork as the reference point is accepted. This dissertation looks at human-machine teamwork at different scales: dyads in Study 1 and Study 2, triads in Study 3, and complex systems in Study 4. Yet, all of these approaches are motivated by the same goal of the current state of the research: this dissertation seeks to identify universal principles of human-machine teamwork that are broadly applicable to a variety of settings.

## 2.3 Reinforcement Learning

### 2.3.1 Background

Reinforcement learning (RL) refers to a class of machine learning algorithms based on behavioral models where behaviors are rewarded and punished until a reward-maximizing policy mapping situations to actions is discovered [Tuyls and Weiss, 2012]. Furthermore, RL agents have to balance exploration and exploitation: exploration can lead to the discovery of new actions and better values over time, while exploitation leads to the selection of the best action available at the time to maximize rewards [Tuyls and Weiss, 2012]. Such tradeoffs are optimized through a series of hyper-parameters and algorithms unique to each RL model.

In recent years RL has gained attention for its performance in Go, Chess, soccer, and Atari games [Hu et al., 1998, Foerster et al., 2018]. In many of these scenarios, such as GO and Dota2, RL proved its usefulness in environments where self-play is possible or where knowledge about other agents is not as important. In both cases, the complexity of the resulting behavior far exceeded that of the environment, and self-play produced a perfectly tuned curriculum for each task [Bansal et al., 2017]. Specifically, DeepMind's research has shown that Deep Q-Networks (combining convolutional neural networks for feature representation with Q-learning training) can achieve superhuman performance in Atari games despite only having access to board states and reward signals [Tampuu et al., 2017].

### 2.3.2 Game Theory and Reinforcement Learning as the Foundation for human-machine Multi-Agent Systems

Game theory is the study of the decision-making behavior of reward-maximizing agents in strategic situations [Von Neumann et al., 2007]. It integrates economics and math to provide a framework that encodes incentive structures in a way that can be understood by both the human (through games) and the AI (through matrices). The fundamental theorem of game theory developed by John Nash guarantees that in every game there will be a set of strategies that each player will converge to as they mutually respond to each other: this is known as the Nash Equilibrium (NE) [Nash et al., 1950].

However, it is often the case that the NE results in an outcome that is either suboptimal or collectively harmful to the players, and yet is inescapable because it emerges from each agent not having a strictly better alternative given their expectation of the behavior of the other players [Bab and Brafman, 2008]. Such dilemmas can only be resolved through coordinated and joint strategies, by all the players, that will result in optimal rewards for the group as a whole [de Cote et al., 2006]. Coordination, cooperation, reciprocity, and fairness are always contingent on the players' preferences, which are in turn systematically influenced by the game's incentives, and can be designed to tie rewards to collaborative outcomes [Erev and Roth, 1998].

### 2.3.3 Challenges and Limitations

In situations where multiple NEs exist, learning one NE strategy does not necessarily guarantee that the other players will select the same NE strategy, leading to vastly higher levels of complexity [Hu et al., 1998]. Prior research has shown that in a multi-player setting, RL agents converge towards a NE strategy only through myopic actions, training with specific trials, or Q-learning-optimizing greedy algorithms. Often, the agents end up stuck in a local optimum because of the associated costs of exploration [Hu et al., 1998]. Thus, despite their prowess, RL agents are susceptible to game-theory induced dilemmas.

Two major results point to the game-theoretical deficiencies of RL agents. First, in zero-sum games, RL agents overestimate their future discounted rewards to be near 0.5 when the expectation between two equally skilled players would make it zero [Tampuu et al., 2017]. Secondly, in infinitely repeated Prisoner's dilemmas with discounting, randomly initialised RL agents collapse into independent and uncooperative strategies with high probability, failing to learn to cooperate reliably over time by not considering the learning process of the other players [Foerster et al., 2018].

However, up until recently it was not possible to explore the strategic behavior of such a huge variety of RL models in a game theory context. Specifically, much empirical research about the behavior of RL agents with human players is missing. RL uniquely lends itself to the study of human-machine cooperation because there are multiple game theory scenarios that enable both types of players to work together as a multi-agent system.

### 2.3.4 Relationship to the Dissertation

RL is the most useful machine learning paradigm to study human-machine teamwork with because RL agents solve problems by learning through trial and error without pre-assigned models. This characteristic of independence, alongside their demonstrated useful across a variety of problems, makes them perfectly suitable for studies where the human-AI interaction is two-sided as opposed to one-sided. Multi-agent systems are the closest analogues to teams in pure AI settings, thus the insights we gain from that area are important components to understand machine behavior in human-machine teams. Game theory plays a major role in the cooperative outcomes of multi-agent systems, and provides the perfect interface to encode complex strategic situations and incentive structures in a way that enables humans and RL agents to interact on an even playing field. This insight motivates the methodology employed in Study 1 and Study 2 to precisely visualize how humans and machines change their default strategic behavior when interacting with one another. Game theory especially is an important framework to understand how to design RL agents whose reward structures incentivize cooperation with humans.

## 2.4 Team Cognition

### 2.4.1 Overview

The concept of a team naturally lies at the core of team cognition. As opposed to mere groups, teams are composed of members who have specified roles and responsibilities related to solving and working on complex tasks. As such, even though all teams are groups, not all groups are teams, making teams a form of highly specialized groups [McNeese et al., 2014a]. Salas and colleagues (1992) characterize teams as sets of people who interact dynamically, interdependently, and adaptively toward a common and valued goal/object/mission [Salas et al., 1992].

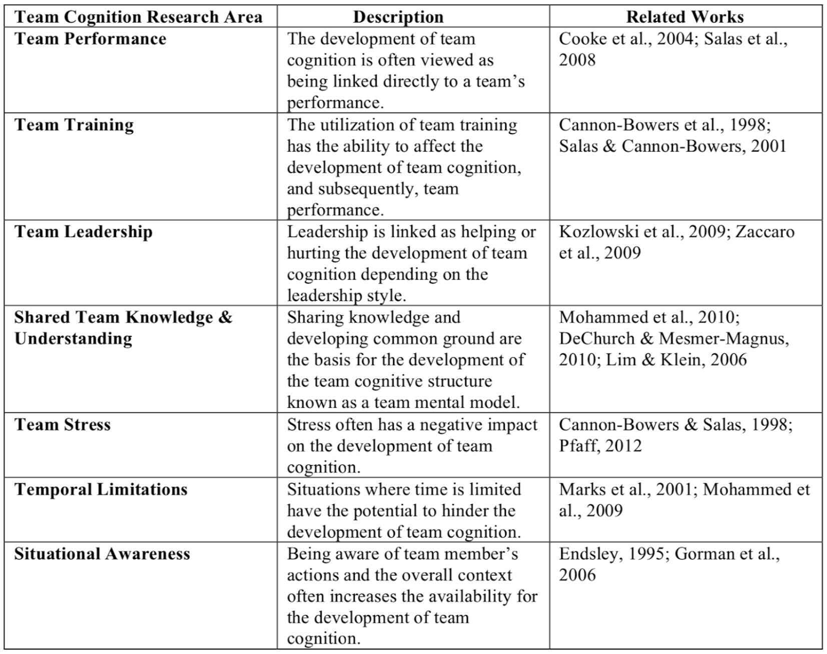
Most of our knowledge pertaining to teamwork is founded on the basis of human- human interactions, including our understanding of situation awareness, teamwork and taskwork knowledge, transactive memory, and team cognition in all-human teams [Cooke et al., 2012].

### 2.4.2 Human Teamwork

At the heart of team effectiveness is communication, for it's the foundation of coordination towards the common objective. Variations on how each team communicates across interactions in tasks and subtasks are a major predictor of performance [McNeese et al., 2014a]. As the team acquires, shares and processes information, a larger cognitive process occurs, which has led prior researchers to characterize teams as information processing units that allow for the attention, encoding, storage, retrieval, and processing of information [Hinsz et al., 1997].

One of the most important aspects of teamwork is team cognition. Team cognition is the glue that can hold a team together by improving communication, coordination, and awareness of the associated teamwork. In its most simple form, team cognition allows for the development of a shared understanding of both teamwork and taskwork [Mohammed et al., 2010]. Team cognition is both a process (communication and coordination) and an output (shared mental model) [Fiore et al., 2010]. There are two main perspectives that conceptualize team cognition: 1) shared knowledge approach, 2) ecological interaction approach [McNeese and Cooke, 2016]. The shared knowledge approach is defined by an input-process-output paradigm where the input is individual team members' knowledge, the process is the sharing of the knowledge, and the output is shared cognition typically represented in the form of a shared mental model [Mohammed et al., 2010]. The ecological interaction approach states that team cognition is team interaction, thus there is significant focus on communication and coordination at the team level [Cooke et al., 2012]. This approach also explicitly notes that context must be taken into account when considering team cognition.

It's vital to note that team cognition is not simply the sum of the individid-



| Team Cognition Research Area | Description | Related Works |
|---|---|---|
| **Team Performance** | The development of team cognition is often viewed as being linked directly to a team's performance. | Cooke et al., 2004; Salas et al., 2008 |
| **Team Training** | The utilization of team training has the ability to affect the development of team cognition, and subsequently, team performance. | Cannon-Bowers et al., 1998; Salas & Cannon-Bowers, 2001 |
| **Team Leadership** | Leadership is linked as helping or hurting the development of team cognition depending on the leadership style. | Kozlowski et al., 2009; Zaccaro et al., 2009 |
| **Shared Team Knowledge & Understanding** | Sharing knowledge and developing common ground are the basis for the development of the team cognitive structure known as a team mental model. | Mohammed et al., 2010; DeChurch & Mesmer-Magnus, 2010; Lim & Klein, 2006 |
| **Team Stress** | Stress often has a negative impact on the development of team cognition. | Cannon-Bowers & Salas, 1998; Pfaff, 2012 |
| **Temporal Limitations** | Situations where time is limited have the potential to hinder the development of team cognition. | Marks et al., 2001; Mohammed et al., 2009 |
| **Situational Awareness** | Being aware of team member's actions and the overall context often increases the availability for the development of team cognition. | Endsley, 1995; Gorman et al., 2006 |

Figure 2.1: Areas of Team Cognition Research

ual cognitions of the team members, but rather the team-level cognition emerging from the interactions of the team members. Many activities team members carry out are done so independently without team-level interactions, thus team cognition is an emergent phenomenon that develops depending on the situation and context [Cooke et al., 2007].

A major finding in team cognition research is that team cognition directly affects team performance. Researchers have noted on multiple occasions that a lack of or breakdown in team cognition may lead to decreased team performance (Wilson et al., 2007). Similarly, it is found that the development of team cognition improves team performance [Cooke et al., 2007, Cooke et al., 2012].

A summary of the different research areas involved in team cognition research are highlighted in Figure 2.1, which has been repurposed with permission from the author [McNeese et al., 2014b]:

### 2.4.3 Shared Mental Models

A major theory of team cognition is the shared knowledge perspective. This theory is predicated upon shared mental models. Mental models as the mechanisms that enable humans to describe and explain a system's purpose, form, function, and predictable future states [Rouse and Morris, 1986]. Mental models emerge from individual cognition, and much in the same way shared mental models emerge from individuals interacting at the team level. Team mental models are thus an emergent property of team cognition as the team members developed a shared understanding and mental representations of knowledge about the team's environment [Cannon-Bowers et al., 1990, McNeese and Cooke, 2016].

Prior research highlights how compatible, shared mental models lie at the foundation of the ability for experienced teams to coordinate, anticipate, predict, and adapt to both the tasks as well as to each other's needs [Fiore et al., 2010]. Mental models are not merely shared through communication, but also through extended observation. These factors enable each team member to describe, explain, and predict future events at the team level [Graham et al., 2004, Mathieu et al., 2000]. Moreover, team mental models are often evaluated in terms of similarity, which in this context refers to the extent to which a team members' knowledge structure is akin to that of an experts' structure [Hamilton et al., 2010]. Related research has highlighted how social network distance and physical distance are major predictors of team mental model similarity, as well as how high levels of communication and strengths of observation magnify the team's shared understanding of individual responsibility and work habits high levels [Graham et al., 2004].

An alternative perspective on team mental models breaks them down into taskwork and teamwork models. Research under this perspective has shown that both type of mental models are positively related to team performance, and that team processes fully mediate the relationship between team effectiveness and mental model converge [Mathieu et al., 2000]. However, many researchers have rejected the view that there is a single mental model that is shared among teammates, and instead argue that multiple mental models are shared throughout the team decision-making process [Salas et al., 1992, Klimoski and Mohammed, 1994]. Under this view, there are four primary mental models:

1. The *equipment* model: the different types of technology and equipment that the team uses to accomplish tasks.

2. The *task* model: the procedures, environments, and tasks the team perceives.

3. The *team interaction* model: the collection of each team member's assumptions, perceptions, and understanding of their teammates' norms, responsibilities, and interactions.

4. The *team* model: the teammates' represents teammates' understanding of each other's attributes, knowledge, and skills.

Beyond classifying team mental models by their content, the literature outlines two other majors classification criteria for mental models: accuracy and similarity [McNeese et al., 2014b]. The former refers to how precisely the mental model matches the real world, as well as the extent to which the team's knowledge structure mimics that of an expert [Edwards et al., 2006, Hamilton et al., 2010, Webber et al., 2000]. The latter refers shared knowledge understood as the degree to which the mental models of the team members' are consistent with each other or converge without becoming identical [Cannon-Bowers et al., 1990, Mohammed et al., 2010]. This accuracy is useful in determining the strength of a team's mental model.

Outside of mental model classification, prior research has focused on investigating mental model convergence among teammates. Mental model convergence occurs through the continuous interplay between communication and interaction among teammates. As team members collect information and observe their teammates' behavior, their individual mental model evolves into a team mental model, thereby shifting the cognitive focus to the team level [McComb, 2007]. This converge processes occurs through three distinct phases: orientation, differtation, and integration . After these three phases, the team converges towards a shared understanding of the model, which is not static but rather is dynamic as the team cycles through the processes in order to adapt to their environment [McComb et al., 2010].

### 2.4.4 Multi-level theory and Macrocognition

Shared mental models speak to the team's cognitive process in the realm of perception. At the decision-making level, the best-known theory is multi-level theory, which identifies four levels of decision-making that are eventually aggregated once a decision happens at the team level [Hollenbeck et al., 1998, Hancock and Szalma, 2008]:

1. Decision Level: at the ground level, the individual team members acquire information and make local decisions to solve specific problems, yet lack the information to solve global problems.

2. Individual Level: at this level the individual makes recommendations to the leader, which can be degrees away or towards the correct decision for the team.

3. Dyadic Level: at this level, the degree to which a team leader correctly weighs each team member's recommendation to arrive at a decision for the whole team.

4. Team level: at this level, the team's hierarchy is tested as the leader seeks to

optimally use and rely on all team member's lower level of analysis.

The results from 380 individuals arrayed into 95 four- person teams working on a simulated naval command and control task indicated that the constructs specified by this theory accounted for over half of the variance in team performance [Hollenbeck et al., 1998].

The multi-level decision-making process often implies a hierarchy, but this kind of cognitive behavior can also be explained through the lenses of macrocognition, which refers to "the internalized and externalized high-level mental processes employed by teams to create new knowledge during complex, one-of-a-kind, collaborative problem solving" [Letsky et al., 2007]. These team processes create new information that the team then utilizes to solve the problem, and they don't necessarily occur purely through observable communication. Prior research identifies five key stages along with fourteen cognitive processes [Letsky et al., 2007]:

1. Individual Knowledge Building
    - (a) Iterative Information Collection
    - (b) Individual Task Knowledge Development
    - (c) Individual Mental Model Development
2. Team Knowledge Building
    - (a) Pattern recognition and Trend Analysis of Team Mental Model Development
    - (b) Recognition of Expertise
    - (c) Sharing Unique Knowledge
    - (d) Uncertainty Reduction

   (e) Knowledge Interoperability

3. Developing Shared Problem Conceptualization

   (a) Visualization and representation of meaning or Building common ground

   (b) Knowledge sharing and transfer

   (c) Team Shared Understanding

4. Team Consensus Development

   (a) Critical thinking

   (b) Mental simulation

   (c) Intuitive Decision Making

   (d) Iterative Information Collection

   (e) Solution Option Generation

   (f) Storyboarding

   (g) Team Pattern Recognition

   (h) Negotiation of Solution Alternatives

5. Outcome Appraisal

   (a) Feedback structure

   (b) Replanning

   (c) Team Pattern Recognition

These cognitive processes however are not just theoretical constructs useful in understanding how teams operate: they are also observable in the differential performance of optimal teams when compared to sub-optimal ones [Letsky et al., 2007].

However, teams are by no means perfect, and often face observable constraints on their cognitive capacity.

### 2.4.5 Functional Theory

The functional theory of group decision-making focuses on the outcomes of the interactions and processes that make teams effective, and it formalizes a set of assumptions: groups a goal oriented, performance and behavior can be evaluated, group interactions are variable and can be evaluated, and lastly group performance is influenced by both internal and external factors [McNeese et al., 2014b]. Furthermore, it attributes improved group decision-making to the effectiveness of the group as a whole.

Most research from the functional perspective focuses on the relationship between group decision-making and the team's ability to satisfy five primary functions [Orlitzky and Hirokawa, 2001] :

1. *Problem Analysis*: The group must develop a thorough and accurate understanding of the nature of, the criticality of, the likely cause of, and consequences of the problem

2. *Establishment of Evaluation Criteria*: The group must develop standards to define, understand, and evaluate a successful response to the problem among alternatives

3. *Generation of Alternative Solutions*: The group must develop realistic alternative solutions to the problem while operating under the assumption that a correct solution exists.

4. *Evaluation of Positive Consequences of Solutions*: Given that there will be

multiple options to select from, it is critical that the group understand and analyze the positive merits of each developed alternative solution.

5. *Evaluation of Negative Consequences of Solutions*: The group must weigh the negative consequences in the same way it weighs positive consequences in order to work a common understanding.

Under functional theory, performance and group effectiveness are expected to depend on the frequency and quality of the previous five interactions engaged with by the group. A meta-analysis of the empirical results from the functional theory literature underscores that the single most important process to group effectiveness is the team members' ability to assess the negative consequences of alternative solutions [Orlitzky and Hirokawa, 2001] . Thus, groups who fully understand the problem their asked with and effectively evaluate the merits of alternative solutions with a keen eye for negative consequences are predicted to be effective under the functional theory of group decision-making

### 2.4.6 Limitations of Teams

Even though teams adapt and respond to complex situations effectively, they often still fall prey to the biases that affect individuals. For example, groups can be primed to over-emphasize solutions from one problem to subsequent ones. Priming in groups can inhibit creativity to solve complicated problems and cause groups to resemble individuals in terms of mental set or habitual routine [Hinsz et al., 1997].

Furthermore, large teams are often inefficient at storing information. It is estimated that groups use only about 70% of their storage capacity because of the losses incurred from the collaboration required to remember at the group level [Hinsz et al., 1997].

Beyond that, in scenarios marked by deep uncertainty, where probabilistic thinking is key to navigating the situation, teams not only fail to escape the base rate fallacy, as the team members reflect a tendency to neglect base-rate information when making a judgement, but they also exaggerate this very tendency [Hinsz et al., 1997].

The research reviewed above indicates that groups appear to exaggerate the tendencies of information processing that occur among individuals. If some bias, error, or tendency predisposes individuals to process information in a particular way, then groups exaggerate this tendency. However, if the bias, error, or tendency is unlikely among individuals processing the information (e.g., less than half of the sample), then groups are even less likely to process information in this fashion [Hinsz et al., 1997].

### 2.4.7 Relationship to the Dissertation

Teams not only display magnified cognitive capacity, but they also display unique cognitive abilities that emerge from the interaction between the team members. Those abilities however are not unlimited and are often constrained by the very biases that plague individual decision making. However, much in the same way team cognition emerges from individuals to produce intelligence and behaviors beyond the individual, a different type of intelligence emerges from large groups and crowds that cannot be reduced to behaviors of the individual – collective intelligence.

## 2.5 Collective Intelligence

### 2.5.1 Overview

Crowdsourcing efforts have produced remarkable insights, findings, and inventions that are hardly expected by individuals working alone or together. A primary example is Foldit, a crowdsourcing effort for biochemistry and protein folding that uncovered in just three weeks the structure of an enzyme related to AIDS that had eluded scientists for 15 years [Malone, 2018]. This type of phenomenon taps into resources and skills needed to perform an activity are distributed widely or reside in places that are not known in advance [Malone et al., 2009].

Such valuable, productive and intelligent behavior emerges from decentralized groups of people that explore and aggregate local information into collectively useful knowledge [Surowiecki, 2005]. Even though the decentralized nature of collective intelligence stands in sharp contrast to centralized and interdependent nature of team cognition, the two phenomena can be understood as manifestations of the same emergent properties.

### 2.5.2 Background

Collective intelligence is the result of the proper aggregation of local information in generating a global solution to a problem that is more optimal than what any individual could have provided [Watkins, 2007]. However, collective intelligence should not be confused with "groupthink": it is not merely the sum product of group opinions but is instead a weighed and calibrated end-product of an information exchange between a group of thinkers. Just because a group convenes and votes on an issue, it does not mean that the "wisdom of the crowd" is occurring.

Specifically, not all groups are good knowledge generators. At the extreme, a crowd morphs into a mob: a dangerous and efficient arrangement to distribute knowledge to members. Even at a micro-level, teams often fail to integrate all relevant information about a problem before making a decision due to the kind of pressure towards conformity inherent to group interactions. Social norms can pressure individuals with distinct perspectives to alter their behavior in order to assimilate, which undermines the kind of diversity that lies at the core of the accuracy gains in collective intelligence [Watkins, 2007].

Prior research has identified four conditions that enable the emergence of collective intelligence in a crowd [Surowiecki, 2005]:

1. Diversity of opinion: each person should have some private information, even if it's just an eccentric interpretation of the known facts
2. Independence: people's opinions are not determined by the opinions of those around them
3. Decentralization: people are able to specialize and draw on local knowledge
4. Aggregation: some mechanism exists for turning private judgments into a collective decision

Prediction markets succeed because their nature lends itself to support all four factors, as participants have a financial incentive to research and grain private information that is then implicitly shared once they begin trading in the market. Decentralization is especially apparent, as any individual gets to immediately trade with every other participant in the market, enabling information to flow very rapidly because it does not have to go through a hierarchy.

Independence is also extremely important to collective intelligence because the underlying reality of any crowd effort is that no individual has perfect access to all information, and that the estimate of all individuals is always flawed in some way. Independence guarantees that errors in individual judgment won't wreck the group's collective judgment as long as those errors aren't systematically pointing in the same direction. One of the quickest ways to make people's judgments systematically biased is to make them dependent on each other for information. Furthermore, independent individuals are more likely to have new information rather than the same old data everyone is already familiar with [Surowiecki, 2005].

Decentralization at its best ensures in a balance between independence and coordination and between specialization and aggregation. At its worst it fails to guarantee that value info that is present within the system (within an individual or a small sub-group) is going to be propagated throughout the rest of the system [Surowiecki, 2005]. Collective intelligence can thus only emerge if the individuals specialize and acquire local knowledge and the crowd (a market, corporation, agency) aggregates into a globally and collectively useful whole [Surowiecki, 2005].

In essence, the reason why the average of all a classroom's estimates for how many jelly beans are in a jar is only a few percentage points away from the actual number is because the overconfident estimates and underconfident estimates offset each other, thereby distilling signal from the noise and yielding an estimate that is superior to that of any individual participant.

### 2.5.3 Prediction Markets as Collective Intelligence Mechanisms

Prediction markets are mechanisms that enable participants to bet upon the occurrence of particular events. At their core, they extend the dynamics of the stock market, where traders buy and sell stocks in anticipation of corporate announcements, to broader events such as political elections and box office performance. A basic example would be an election, where the value of a candidate's "stock" becomes $1 if the candidate wins, and $0 if the candidate loses, thus enabling participants to buy and sell the stock until all trading ends and yields a price that inherently reflects the probability of the candidate winning (70c would imply a 70% chance).

Prediction markets are remarkably effective at forecasting events and are often better than pundits and experts alike. For instance, in the case of sports, real-money prediction markets were found to be more accurate than expert polls [Goel et al., 2010] . In the realm of politics, a study of the Iowa Electronic Markets (IEM)'s performance over the course of the presidential elections between 1988 and 2000, shows that the IEM's market price on the day each of the 596 different polls were released was more accurate than the polls themselves 75% of the time [Hanson, 2003, Surowiecki, 2005]. These results carry over into geopolitical forecasts as well, where IARPA's DAGGRE prediction market accuracy was about 38% greater than the baseline system at over 400 geopolitical questions [Laskey et al., 2015]. Recent research also suggests that prediction markets outperform even AI-based big data approaches. For instance, IEM outperformed a highly advanced machine learning model analyzing 40 million unique tweets in the 2012 election [Attarwala et al., 2017]. These results are not isolated, but rather are consistent with a broader pattern of prediction markets being systematically more effective than expert and collective judgements.

The key to the success of prediction markets lies in their ability to aggregate diverse opinions to parse signal from noise. Specifically, prediction markets can combine potentially diverse opinions into a single consistent probability distribution (Hanson 2003). Markets also provide strong economic incentives for individuals to correct systematic biases, such as overconfidence or underconfidence. A rational trader would place bets that are profitable in expectation, realigning prices with historical base rates [Atanasov et al., 2016]. Furthermore, they are able to handle more complexity than an individual or centralized body could grasp because "knowledge that is implicit, dispersed, and inaccessible by traditional, conscious methods can be organized through markets to create more rational calculation than can elite experts"

[Watkins, 2007].

Watkins (2007) summarizes the three main theoretical explanations for the collective intelligence exhibited by prediction markets:

1. Crowds are not always necessarily better at solving a problem than an individual, yet they are structurally well positioned to overcome the limited informational capacity of individual humans. Thus, as long as each member of the group does not shift their opinion in order to conform to the perceived consensus (groupthink), a prediction market will be even better than its individual members at processing information.

2. Prediction markets effectively balance the synthesis of large amounts of information with the avoidance of interactions that can lead to groupthink. The competitive dynamics of prediction markets do this by encouraging competition as opposed to consensus, thereby creating a strong incentive to avoid sharing privately held information that can influence of pressure others into ceasing independent decision-making.

3. Prediction markets ensure high degrees of the type of diversity critical to decision-making by disincentivizing copycat behavior. The financial stakes required to participate in a prediction market thus function as a self-selection mechanism that disincentivizes participation by individuals who do not have marginal information to contribute to the market. Specifically, the incentives of a prediction market force individuals to assess the uniqueness of information they possess as they attempt at profiting from the market price adjusting to reflect their expectations.

Wolfer (2009) advances another theory as to why prediction markets are so effective: the marginal trader hypothesis. The hypothesis states that *"the efficiency* *of prediction markets is driven by a minority of unbiased and active participants who wield corrective influence"*. This view argues that collective intelligence emerges in prediction markets because the mechanism incorporates differences in forecaster knowledge and skill automatically. Specifically, in the short turn order size serves as a useful proxy for a trader's confidence in a prediction, and in the long run higher earnings are rewarded to traders who made correct predictions, thereby increasing the resources at their disposal to influence future prices. What drives trader behavior is thus the rational and risk-averse expectation to profit from the difference between the market price and their private beliefs [Wolfers, 2009].

In essence, prediction markets generate a strong financial incentive for participants to express their opinions in precise and informative ways, which are then aggregated and calibrated by the trading mechanism to reflect the most reliable estimate for a particular event. This transition from the local information of each participant to the global information of the crowd speaks to the powerful emergent property of prediction markets: collective intelligence.

### 2.5.4 Limitations of Prediction Markets

Just like financial markets, prediction markets are not immune to problems. Forecasting future events is such a challenging task that is prone to errors of all types, that can potentially be magnified by the macro-nature of a prediction markets. Prior research has identified three major types of errors in prediction markets [Dudik et al., 2017]:

1. Sampling error, which arises from traders possessing noisy estimates that dilute the truth-value of their information.

2. Market-maker bias, arising from a particular cost function being used to gen-

erate an opportunity for profit to facilitate trading actually inducing particular biases on overshooting or undershooting the estimate.

3. Convergence error, arising from huge market fluctuations caused by all the trading before the price stabilizes arising because, at any point in time, market prices may still be in flux

These problems are exacerbated by the fact that redesigning the aggregation function, the primary technical solution often discussed in the literature, often not enough. Chen et al. (2005) analyzed data from football games and found that linear, logarithmic, absolute distance, and quadratic scoring did not differ significantly as aggregation functions in their overall accuracy. Purely technical approaches have thus not shown major improvements in addressing some of the variability factors in prediction markets.

An alternative that has been proposed is to move away from prediction markets all together and focus on better prediction polls to elicit and aggregate estimates from individuals. The results however have not been supportive of such a claim, as prior research has indicated that simple aggregate of prediction polls tends towards not just underconfidence (despite the well-known tendency for people to be overconfident) but also less meaningful as the average forecasts converge towards 50% probability for two-option questions [Satopää et al., 2014].

Lastly, prediction markets rely on financial incentives to motivate participants. This makes sense, for no participant would trade if there was not an opportunity to profit from someone else's lack of information (Ex. buying a candidate's stock that sells for 50c to resell it at 70c). Thus, prior researchers have highlighted the importance for the manager of the prediction market to subsidize trading for new events in order to catalyze the trading process among participants

[Chen and Vaughan, 2010, Chen et al., 2010, Hanson, 2003]. This kind of solution creates a barrier to entry to the implementation of prediction markets in areas beyond geopolitics or elections, where a thick market of many participants can be expected.

We thus seek to address these concerns by designing a new type of prediction market that not only harnesses collective intelligence in a sustainable way, but also enables artificial intelligence to address many of the limitations of traditional design.

### 2.5.5 Relationship to the Dissertation

Collective intelligence provides an alternative perspective with which the emergent cognitive properties of human-machine teams can be understood. However, the role AI can play in prediction markets is still very poorly understood, and this dissertation seeks to remedy that through Proposed Study 2. Studying the applications of machine learning to prediction markets will serve as useful methodology to determine whether the biases identified in this section (biases the literature considers inherent to the cognition of the participants) can be mitigated by treating prediction markets as a large-scale human-machine team. Moreover, many of the principles of collective intelligence stand in sharp contrast to those of team cognition, thus reconciling both will inform a much better way to understand human-machine teamwork by determining which principles from either model are applicable.

## 2.6 Conclusion

This dissertation aims at developing a comprehensive view of human-machine teamwork that accounts for the different role AI can play in a human-machine team. To that end, this chapter highlighted the relationship between team cognition and collective intelligence: they are all emergent cognitive phenomena that enables teams to accomplish tasks that are beyond the grasp of the individual. The next chapter presents a study we conducted focusing on the role game theory plays in the cooperation between humans and RL agents.

# Chapter 3

# Study 1: Game Theory for Teams

A major theme of the dissertation is understanding human-machine teamwork as a two-sided relationship. In practice, this means approaching human-machine teamwork from both the perspective of the human as well as the perspective of the AI. The following study of human-AI cooperation in several game theory scenarios does this by first looking at how different RL models alter their cooperative behavior when interacting with humans; and secondly by looking at how the human participant's belief about whether they are playing against a human or an AI influences their willingness to cooperate. The key findings for this study speak to RQ1 by showing how cooperation can occur between humans and machines even when communication is not possible, while also proving that incentive structures play a major influence in human-machine teamwork.

Reported here are major excerpts of the journal article that was subsequently submitted and is now under review.

## 3.1 Introduction

The current trajectory of the field of computer science is aimed at developing artificial general intelligence (also known as AGI). This kind of strong AI would theoretically be capable of matching and eventually outperforming human intelligence across a variety of different tasks where intelligence plays a decisive role [Goertzel and Pennachin, 2007]. AGI differs from the kind of narrow AI that we see today in that, unlike current AI, it can easily transfer learning between tasks because of a true understanding of itself and its environment [Pennachin and Goertzel, 2007]. Regardless of whether AGI will ever be feasible, the focus of AI research should not be towards developing technology meant to make humans obsolete. Instead, the AI community should reconsider whether the inevitable consequence of the rise of AI will be the full replacement of human effort. In fact, prior research suggests that effective human-AI partnerships outperform not just human teams, but also sophisticated AI systems.

A major historical example of the valuable yet counter-intuitive nature of the findings that emerge from human-machine teamwork research is in chess, with IBM's Deep Blue, a chess-playing computer that eventually defeated Gary Kasparov, the world champion at the time. Kasparov had already won the first match against Deep Blue, and it was only after a substantial hardware upgrade that IBM's Deep Blue finally defeated Kasparov [Hsu, 2004] . Deep Blue however did not really prove computers superior to the human brain at a complex task such as chess; rather, its performance was driven by a brute force evaluation of every possible move, as opposed to heuristic and strategic approach that Kasparov and any chess player engages in [Cords, 2007]. This key difference gets precisely at one of the primary distinctions between artificial and human intelligence: the human brain is not merely running computations, but rather is making sense of the larger context of the situation [Hipp et al., 2011].

Kasparov went on to demonstrate through his "advanced chess" tournament (where AIs, humans, and human-machine teams compete against each other), that a human-machine team could defeat both the top AI as well as the top human chess players, and that the human-machine team was not constituted of a partnership between a top chess player and a sophisticated AI. Rather, an amateur human and a mediocre AI managed to outperform precisely because their limitations made their intelligence level more compatible as the human focused on highlighting the top moves they were considering and the AI computed opponent responses to figure out the best move among those options [Thompson, 2010]. This type of finding suggests that through collaborative interfaces that bridge the gap between AI and the human brain, a shared understanding of the situation can emerge that enables human-machine teams to perform at their best.

More recently however, the team at DeepMind managed to create AlphaGo: an AI that defeated the world champion in Go [Silver and Hassabis, 2016]. AlphaGo is distinct from Deep Blue because it's dealing with a much more complex problem: the number of possible moves in Go is exponentially greater than that of Chess, thereby making it impossible for the AI to compute all possibilities [Schraudolph et al., 1994]. Through machine learning, AlphaGo, unlike Deep Blue, developed a strategic understanding of its task, thereby discovering brand new ways to play the game that were unknown to humanity beforehand. Furthermore, once upgraded, AlphaGo achieved mastery of the game by playing against itself as opposed to training its analytical skills with games from famous players [Silver et al., 2017]. However, it is easy to draw the wrong lesson from this event. Despite its heightened capabilities, AlphaGo is still a narrow AI whose usefulness is strictly limited to singular, well-defined tasks and thus cannot adapt to context changes or complex effects.

Naturally, it becomes paramount to delve more deeply into the mechanics of a successful human-machine partnership. To start with, it's useful to extend the constructs of human-human teaming to ground our understanding of human-machine teams. Teams at their core are composed of interdependent agents who at their highest level operate according to a shared understanding of the task and situation they are confronted with [McNeese et al., 2017]. Through this shared situational awareness, teams can adapt to a dynamic environment while retaining coordinated behavior that's critical to accomplish both short-term and long-term goals [McNeese et al., 2017]. Communication is what enables individuals to relate to and understand one another to a sufficient enough extent that they begin processing information cohesively, which leads to the emergence of team cognition [Demir et al., 2016, Demir et al., 2017].

The challenge with human-machine teams however is that both types of agents operate with fundamentally different understandings of the world that cannot be easily communicated. Prior research has shown that effective team behavior occurs when each team member seeks to model the thought process of their teammates, which is inherently more challenging in a human-machine team [McNeese et al., 2017]. Specifically, humans tend to inherently distance themselves from teammates they perceive to be autonomous, and AIs tend to avoid wanting to cooperate with human agents who don't share their thought process [Demir et al., 2018]. These challenges require a bidirectional solution that emphasizes both the need for better thought-sharing protocols as well as better computational architectures that enable the AI to model the human teammate's thought-process [Chattopadhyay et al., 2017]. To achieve the same result from the human-side, it also becomes necessary to simplify the complexity of the AI-agent in a way that is accessible to the teammates and encourages a shared understanding and shared cognition occurring in a human-machine team [Crowder and Carbone, 2014].

It thus follows that further research is needed to clarify the dynamics behind human-machine teams as well as human-machine multi-agent systems as they become more prevalent.

## 3.2 Methods

We enlisted over 600 participants from Amazon Mechanical Turk to play a different game theory scenario under different conditions. The participants were grouped in batches of 50, and were assigned to one of the two possible games. The participants were either told they were playing against an AI or against a human, and were assigned to one of two possible games. We also collected demographics and survey of team effectiveness and perceptions but we are not including those results because they are beyond the scope of this study. The demographics breakdown is displayed in Table 3.1.

| | Male | Female |
|---|---|---|
| Gender | 23.77% | 76.23% |

| | African American | Asian | Caucasian | Indian | Other |
|---|---|---|---|---|---|
| Race & Ethnicity | 3.52% | 37.89% | 40.08% | 16.75% | 1.76% |

| | 18-25 | 26-35 | 36-45 | 46-55 | 56-65 |
|---|---|---|---|---|---|
| Age | 25.55% | 53.73% | 10.13% | 6.16% | 4.40% |

Table 3.1: Demographics

We built an interface that supports all of the experimental conditions, thereby restricting each participant to only interact with a specific variation of the experimental setup. Each move made by the players is recorded by the application and stored on a server. Each player plays 3 rounds of 10 turns each, with the score resetting every round.

To implement the RL agents, we used TensorForce which is an open source RL library focused on providing clear APIs, readability and modularization to deploy RL solutions both in research and practice [Schaarschmidt et al., 2017, Schaarschmidt et al., 2018]. There are a host of predefined algorithms present in this library. The following is a list of algorithms available:

- A3C using distributed TensorFlow [Mnih et al., 2016].
- Trust Region Policy Optimization [Schulman et al., 2015].
- Normalized Advantage functions (NAFs) [Gu et al., 2016].
- DQN [Mnih et al., 2013].
- Double-DQN [Van Hasselt et al., 2016].
- Vanilla Policy Gradients (VPG/ REINFORCE) [Williams, 1992].
- Deep Q-learning from Demonstration (DQFD) [Hester et al., 2018].
- Proximal Policy Optimization (PPO) [Schulman et al., 2017]

We will be focusing on PPO, DQN and VPG for our experiment and results. This methodology led to a 2x2x3 experimental design for a total of 12 experimental conditions. All of the experimental conditions are identified in Table 3.2

## 3.3 Data Collection and Analysis

The data was collected by our game interface automatically. Each time a human player submitted a move, the client sent a request to the REST API in our

| Prisoner's Dilemma | | | |
|---|---|---|---|
| | PPO | DQN | VPG |
| Human Opponent | pd-human-ppo | pd-human-dqn | pd-human-vpg |
| AI Opponent | pd-ai-ppo | pd-ai-dqn | pd-ai-vpg |

| Battle of the Sexes | | | |
|---|---|---|---|
| | PPO | DQN | VPG |
| Human Opponent | bos-human-ppo | bos-human-dqn | bos-human-vpg |
| AI Opponent | bos-ai-ppo | bos-ai-dqn | bos-ai-vpg |

Table 3.2: 2x2x3 Experimental Design

server. The requested triggered a call to the specific RL agent which prompted the agent to submit their own action. Subsequently, the server processed the human and AI move and responded with the appropriate payoffs. Throughout this process, each turn was logged into our database for further analysis.

Specifically, the data collected was the userID, the human player's move (0 for cooperation, 1 for defection), the RL agent's move (0 for cooperation, 1 for defection), the type of RL agent (”ppo”, ”vpg”, ”dqn”), the belief condition (whether the human believed they were playing against a human or AI), and the turn number.

Once the experiment was completed, several R scripts were deployed to clean up and restructure the data so that it could be analyzed properly. Specifically, each pair of moves was translated into an outcome (cooperation, non-cooperation), and each move by the human as well as the RL agent was also classified as cooperative or non-cooperative. The latter data was used to produce histograms plotting the frequency of specific behaviors by both the human and the AI.

Next, given the non-normal nature of the data, we deployed a generalized linear model with mixed effects. Linear mixed models differ from mere linear models in that they are not described by a distribution of vector-valued random response, but rather by the distribution of two vector-valued variables: the response variable and the vector of random effects [Bates et al., 2014]. The term ”mixed effects” refers to the models' incorporation of both fixed and random effects in its linear predictor of the conditional mean of the response variable [Bates et al., 2014]. The model works because the transformation of our collected data turns the 4 possible outcomes of each game into a binomial (cooperation vs non-cooperation).

## 3.4 Game Theory Scenarios

Two cooperative game theory scenarios were selected to provide a broad analytical base to identify the extent to which different factors affect the willingness to cooperate of both the human players as well as the reinforcement learning agents.

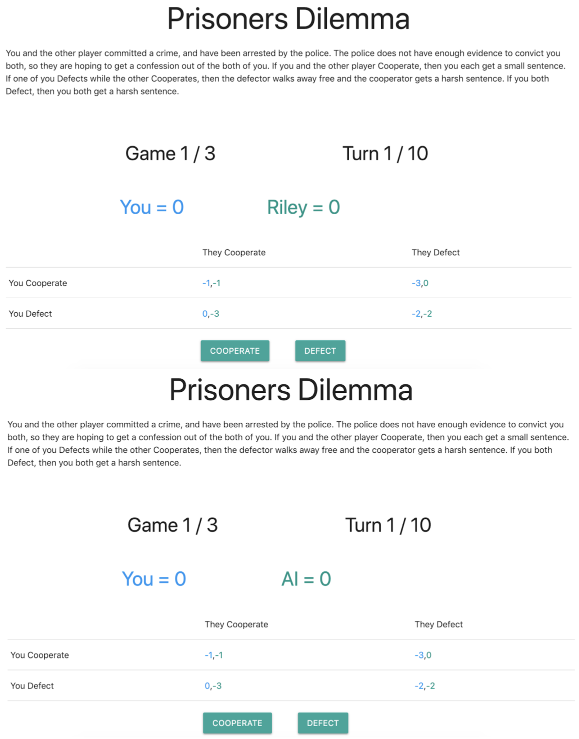
### 3.4.1 Prisoner's Dilemma

The Prisoner's Dilemma (PD) is a classical scenario in game theory where two players are posited to have been arrested by authorities for committing a crime. Once apprehended, each player is separated from the other so that they are unable to communicate. Because the police do not have sufficient evidence to convict both players, they offer to each player the opportunity to confess in order to gain a lighter sentence at the expense of the other player.

The core result in PDs is that the Nash Equilibrium induces both players to confess, leading to the collectively worst outcome for both players. However, in experimental settings this dynamic often changes when the PD is played in an iterative fashion. This is because a sequential PD creates the opportunity for players to punish one another for defecting from an agreement to remain silent, thus creating a reasonable expectation of cooperation.

Figure 3.1 is taken from the PD interface of our web application. The payoffs for mutual cooperation are -1 for each player, the payoffs for mutual defection are -2 for each player, and the payoff for successfully defecting on a cooperative player are 0 and -3 respectively. In a PD, defecting is the dominant strategy because both players are better off defecting given what they expect the other player to do. Essentially, the mutual best response in the game is for both players to defect on one another.

Figure 3.1: Human (bottom) and AI(top) views of the Prisoner's Dilemma Interface



Prisoners Dilemma

You and the other player committed a crime, and have been arrested by the police. The police does not have enough evidence to convict you both, so they are hoping to get a confession out of the both of you. If you and the other player Cooperate, then you each get a small sentence. If one of you Defects while the other Cooperates, then the defector walks away free and the cooperator gets a harsh sentence. If you both Defect, then you both get a harsh sentence.

Game 1 / 3 Turn 1 / 10

You = 0 Riley = 0

| | They Cooperate | They Defect |
|---|---|---|
| You Cooperate | -1,-1 | -3,0 |
| You Defect | 0,-3 | -2,-2 |

Prisoners Dilemma

You and the other player committed a crime, and have been arrested by the police. The police does not have enough evidence to convict you both, so they are hoping to get a confession out of the both of you. If you and the other player Cooperate, then you each get a small sentence. If one of you Defects while the other Cooperates, then the defector walks away free and the cooperator gets a harsh sentence. If you both Defect, then you both get a harsh sentence.

Game 1 / 3 Turn 1 / 10

You = 0 AI = 0

| | They Cooperate | They Defect |
|---|---|---|
| You Cooperate | -1,-1 | -3,0 |
| You Defect | 0,-3 | -2,-2 |

### 3.4.2 Battle of the Sexes

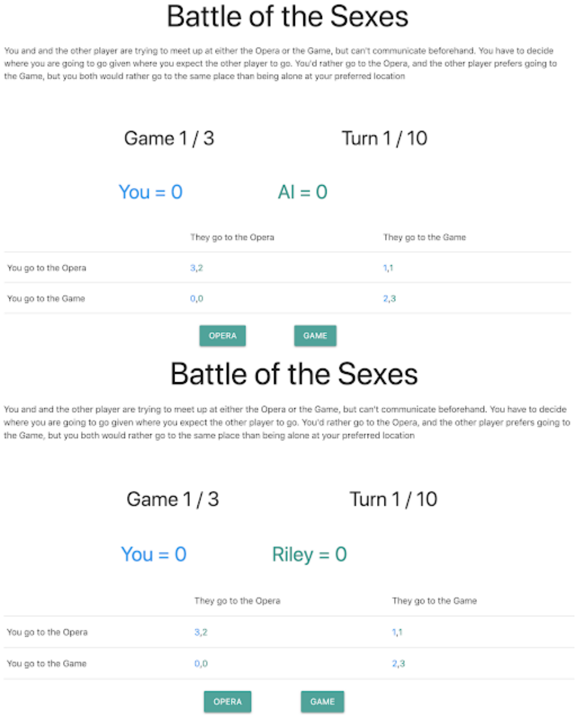
The Battles of the Sexes (BoS) is another classical game theory scenario which posits that two players are trying to meet up at one of two locations. Each player has a preference for which one of the two locations they would rather meet up in, yet both players would rather be in the same place together than by themselves in their respectively preferred locations. The challenge of a BoS is that the players cannot communicate, and thus have to rely on the incentive structure of the game to successfully cooperate.

Unlike with a PD, a BoS does not have a strictly dominant strategy any of the players is better of adopting irrespective of what the other player does. There are three Nash Equilibria: the two pure equilibria associated with each of the two options as long as both players select the same one, and the mixed strategy equilibrium of the players randomizing between the two locations given a calibrated probability function that can be derived from the differential payoffs of the game.

Figure 3.2 is taken from the BoS interface of our web application. The players are given two options: going to Opera or going to the Sports Game. The human player is assigned the preference for the Opera, whereas the AI is assigned the preference for the Game. The human player and the AI earn 3 and 2 points respectively if they both decide on the Opera, and 2 and 3 points respectively if they settle for the Game. Each player also incurs a payoff of 1 point if they fail to end up in the same location are at least in their respectively preferred location, as opposed to 0 points in the worst case scenario where both players not only fail to be in the same location but are also away from their preferred option.

An interesting aspect of a BoS is that it involves a deeply strategic evaluation of whether altruism pays off, because both players selecting the option the other player prefers as opposed to their own results in the worst possible outcome. Thus, the expectation for altruism has to be unilateral. Furthermore, altruistic behavior in this case requires one player to be submissive to preferences of the other player because once the player settles on the same location, the dynamic of the Nash Equilibrium reinforces the incentive not to deviate.

Figure 3.2: Human (bottom) and AI(top) views of the Battle of the Sexes Interface



Battle of the Sexes

You and and the other player are trying to meet up at either the Opera or the Game, but can't communicate beforehand. You have to decide where you are going to go given where you expect the other player to go. You'd rather go to the Opera, and the other player prefers going to the Game, but you both would rather go to the same place than being alone at your preferred location

Game 1 / 3 Turn 1 / 10

You = 0 AI = 0

| | They go to the Opera | They go to the Game |
|---|---|---|
| You go to the Opera | 3,2 | 1,1 |
| You go to the Game | 0,0 | 2,3 |

Battle of the Sexes

You and and the other player are trying to meet up at either the Opera or the Game, but can't communicate beforehand. You have to decide where you are going to go given where you expect the other player to go. You'd rather go to the Opera, and the other player prefers going to the Game, but you both would rather go to the same place than being alone at your preferred location

Game 1 / 3 Turn 1 / 10

You = 0 Riley = 0

| | They go to the Opera | They go to the Game |
|---|---|---|
| You go to the Opera | 3,2 | 1,1 |
| You go to the Game | 0,0 | 2,3 |

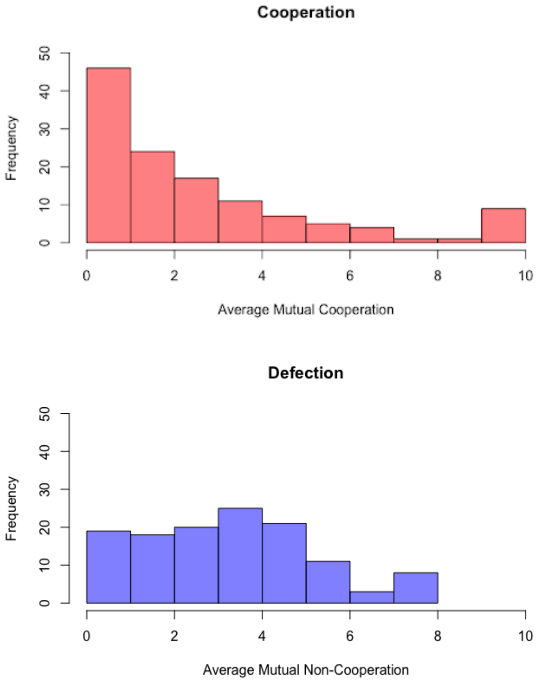
## 3.5 Results

### 3.5.1 Prisoner's Dilemma

Figure 3.3 below shows the frequency distributions of the average mutual cooperation and mutual defection levels between the human and AI participants. The right skew on the first graph indicates that the human and the AI failed to effectively cooperate most of the time to avoid losing points. The second graph shows that on average, humans and AIs did not necessarily spend most of the turns mutually betraying one another.

The distributions make it very clear that the data is not normally distributed, which is to be expected given how the game has a Nash Equilibrium where the players are expected to converge. Therefore, we converted the data to a binomial distribution of cooperation vs. non-cooperation where cooperation represented the

Figure 3.3: Mutual Cooperation and Mutual Defection Frequency Distributions



mutually beneficial outcome of the game. That enabled us to run a generalized linear model with mixed effects:

$$cooperation = opponent + turn + (1 \mid participantID))$$

Table 3.3 shows that the random intercept of the participant did not have residuals, thus suggesting that there is no variability that is not accounted by the random effects of the model. On the fixed effects, the slopes of the different agent type had around similar negative coefficients (-1.41 and -1.53) on overall cooperation that were both significant (p-value = 0.00589 and p-value = 0.00407). The coefficient for turn (which refers to which turn out of the 30 played is being analyzed) is quite small, having a much smaller effect (-0.221) that albeit significant indicates that the levels of cooperation between the human and different AI models are better predicted by which AI model the human playing against as opposed to how long the two players have been playing together.

| Opponent + Turn | | | |
|---|---|---|---|
| | Estimate | Std. Error | p-value |
| **Fixed Parts** | | | |
| (Intercept) | 0.16721 | 0.34310 | 0.62600 |
| opponent:ppo | -1.41327 | 0.51318 | 0.00589 |
| oppponent:vpg | -1.53198 | 0.53327 | 0.00407 |
| turn | -0.22197 | 0.01551 | <2e-16 |
| **Random Effects** | | | |
| | Variance | Std. Dev | Count |
| humanID | 5.281 | 2.298 | 125 |
| Observations | | | 4830 |

Table 3.3: Generalized Linear Model with Mixed Effects for Prisoners Dilemma

We also ran a similar model to determine whether there were any meaningful differences on human behavior if the human was told they were playing against an AI or not. Specifically, we compared a model considering only the belief condition (whether the human player thought they were playing against another human or an AI) and the amount of turns elapsed to a model which also took into account the specific RL model the human was playing against:

$$cooperation = condition + turn + (1 \mid participantID))$$

$$cooperation = condition + opponent + turn + (1 \mid participantID))$$

Table 3.4 shows that the human player's belief that they are playing against another human has a negative effect (-0.73, p-value = 0.04) on overall cooperation. Just like in the prior setup, the number of turns elapsed has a minimal yet significant effect (-0.18, p-value $<2e^{-16}$).

|  | Condition + Turn |  |  | Condition + Opponent + Turn |  |  |
|---|---|---|---|---|---|---|
|  | Estimate | Std. Error | p-value | Estimate | Std. Error | p-value |
| **Fixed Parts** |  |  |  |  |  |  |
| (Intercept) | -0.80757 | 0.19709 | 4.18e-05 | 0.01201 | 0.28983 | 0.96696 |
| opponent:ppo |  |  |  | -1.32838 | 0.43397 | 0.00221 |
| oppponent:vpg |  |  |  | -1.40403 | -0.45067 | 0.00184 |
| turn | -0.18071 | 0.01293 | <2e-16 | -0.18084 | 0.012477 | <2e-16 |
| condition:human | -0.73069 | 0.35581 | 0.04 | -0.20575 | 0.43633 | 0.63725 |
| **Random Effects** |  |  |  |  |  |  |
|  | Variance | Std. Dev | Count | Variance | Std. Dev | Count |
| humanID | 4.037 | 2.009 | 175 | 3.703 | 1.924 | 175 |
| Observations |  |  | 6540 |  |  | 6540 |

Table 3.4: Generalized Linear Model with Mixed Effects Comparison

Furthermore, the condition variable in the model is only significant in the first model, for the second's predictive capacity is dominated by the opponent variable.

### 3.5.2 Battle of the Sexes

Figures 3.4 and 3.5 shows the frequency distributions of the average successful cooperation as well as the average failure to coordinate between the human and AI participants. They both separate the different types of coordination: human-preferred coordination means both the human and the AI decided to go to the Opera, which resulted in higher points for the human, while AI-preferred coordination refers to the scenario when the human and the AI both selected to go the Game, which resulted in higher points for the AI.

Both charts display strong left skews which implies that extreme cases did not happen frequently, suggesting that most of the time the strategic interaction converged to benefit the human.

Much like with the Prisoner's Dilemma, the distributions make it very clear that the data is not normally distributed. Therefore, we converted the data to a

Figure 3.4: Frequency Distributions of Successful Cooperation

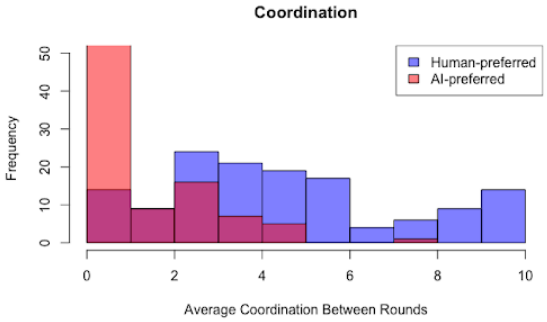
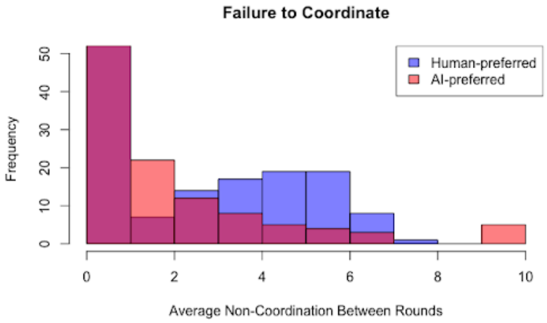


Figure 3.5: Frequency Distributions of Coordination Failure



binomial distribution of cooperation vs. non-cooperation where cooperation represented any of the mutually beneficial outcomes of the game (Opera, Opera and Game, Game). This enabled us to run a generalized linear model with mixed effects:

$$cooperation = opponent + turn + (1 \mid participantID))$$

Table 3.5 shows that the random intercept of the participant did not have residuals, thus suggesting that there is no variability that is not accounted by the random effects of the model. On the fixed effects, the slopes of the different agent type had around similar negative coefficients (-1.36 and -1.10) on overall cooperation that were both significant (p-value $= 5.15e^{-7}$ and p-value $= 7.48e^{-6}$). The coefficient for turn is once again quite small, having a drastically smaller effect (-0.03422) that, albeit significant, indicates that the levels of cooperation between the human and different AI models are better predicted by which AI model the human playing against as opposed to how long the two players have been playing together.

| Opponent + Turn | | | |
|---|---|---|---|
| | Estimate | Std. Error | p-value |
| **Fixed Parts** | | | |
| (Intercept) | 0.97945 | 0.18582 | 1.36e-07 |
| opponent:ppo | -1.36189 | 0.27126 | 5.15e-07 |
| oppponent:vpg | -1.10211 | 0.24603 | 7.48e-06 |
| turn | 0.03422 | 0.01110 | 0.00205 |
| **Random Effects** | | | |
| | Variance | Std. Dev | Count |
| humanID | 1.318 | 1.148 | 137 |
| Observations | | | 4890 |

Table 3.5: Generalized Linear Model with Mixed Effects for Battle of the Sexes

We also ran the same model to determine whether there were any meaningful differences on human behavior if the human was told they were playing against an AI or not. Just like with the Prisoner's Dilemma, we compared a model considering only the belief condition (whether the human player thought they were playing against another human or an AI) and the amount of turns elapsed to a model which also took into account the specific RL model the human was playing against as well:

$$cooperation = condition + turn + (1|participantID))$$

$$cooperation = condition + opponent + turn + (1|participantID))$$

Table 3.6 shows that the human player's belief that they are playing against another human has a slight negative effect (-0.54, p-value=0.0275) on overall cooperation. Just like in the prior setup, the number of turns elapsed has a minimal yet positive effect (0.02, p-value = 0.0190).

| | Condition + Turn | | | Condition + Opponent + Turn | | |
|---|---|---|---|---|---|---|
| | Estimate | Std. Error | p-value | Estimate | Std. Error | p-value |
| **Fixed Parts** | | | | | | |
| (Intercept) | 0.23744 | 0.11333 | 0.0362 | 0.98986 | 0.16478 | 1.89e-09 |
| opponent:ppo | | | | -1.32168 | 0.23967 | 3.50e-08 |
| oppponent:vpg | | | | -1.06341 | 0.21745 | 1.01e-06 |
| turn | 0.02333 | 0.00995 | 0.0190 | 0.02848 | 0.26698 | 0.9151 |
| condition:human | -0.54203 | 0.24591 | 0.0275 | 0.02848 | 0.26698 | 0.9151 |
| **Random Effects** | | | | | | |
| | Variance | Std. Dev | Count | Variance | Std. Dev | Count |
| humanID | 1.301 | 1.141 | 166 | 0.9973 | 0.9987 | 166 |
| Observations | | | 5880 | | | 5880 |

Table 3.6: Generalized Linear Model with Mixed Effects Comparison

The second model has much lower levels of variance accounted for by random effects (0.9973). The result is similar to that of the Prisoner's Dilemma in that the more complete model does not hold the belief condition (whether the human believes the other player to be an AI or not) as statistically significant because most of its predictive power emerges from the type of RL model involved in the interaction.

### 3.5.3 Summary

The generalized linear models with mixed effects proved very effective as predictors of cooperation between the humans and the AIs over a variety of conditions. The amount of variance accounted for by random intercepts was higher in the PD game than in the BoS game, but they were both much more significantly predicted by the fixed effects. Specifically, although the turn number was a significant predictor of cooperation (which implies that the amount of turns the human and the AI have played with each other affects their overall tendency to cooperate), the major predictors were the types of RL model. The PD experiment produced data about cooperation (willingness to sacrifice personal gain for collective gain) and the BoS produced data about coordination (willingness to consider the other player's preferences to achieve an optimal outcome). Lastly, although the belief condition (whether the human believes they are playing against AI or against another human) has a negative influence on the human player's willingness to cooperate, it's only significant when it's used as a predictor along with turn number. The adding more fixed effects to our model did not improve prediction in a statistically significant ways, whereas exploring different combinations of fixed effects did.

## 3.6 Discussion

Our results show meaningful differences in the cooperative dynamics between humans and AIs across a variety of different settings. Instead of limiting ourselves to just one game theory model, such as the often-used Prisoner's dilemma, we explored human-machine cooperation in even more complex social dilemmas. The results from each game can be analyzed separately, but they should also be understood as indicative of a broader pattern of human behavior.

Specifically, the data from the Prisoner's dilemma setup have implications for both AI as well as human cooperative dynamics. For the former, each RL model cooperated differently with their human counterpart. This result is not surprising given the distinct ways in which DQNs and PPOs process strategic interactions. On the other hand, our GLMER also suggests that humans have a higher propensity towards cooperation when playing against other humans as opposed to AIs, which belies a particular bias in the context of a Prisoner's dilemma. Furthermore, the results from the Prisoner's Dilemma dynamic stand out as strong indications that both humans and AIs hold hostile tendencies in such a context. At the outset, the

Nash Equilibrium of the game induces both players towards aggression. However, the iterative nature of our setup would lend itself towards the development of norms of reciprocation, where both players would learn over time that aggression would be punished and thus potentially converge towards peaceful outcomes. The GLMER provides similar support for cooperative outcomes being mediated by the type of RL agent involved (hypothesis 1) as well as to whether the human believes they are playing an AI or not (hypothesis 2).

The results from the Battle of the Sexes should also be analyzed separately. To start with, the game itself does not have just one Nash Equilibrium, giving the players two possible outcomes where they mutually benefit from cooperation. The data indicates that the most frequent outcome was cooperative and to the benefit of the human (I.E. the players converged towards the Nash Equilibrium that benefits the human most). The GLMER shows that although the type of RL agent strongly influence the outcome of the interaction (hypothesis 3), the nature of the equilibria leads the players to cooperate more than in the two prior instances.

Overall, larger patterns about the dynamics of human-machine cooperation can be observed across all three games. First, not a single game resulted in the players converging towards the Nash Equilibrium 100% of the time. This is important because it speaks to limitations of theoretical research on human-machine teamwork. Only through empirical setups, such as ours, can the belief that human-machine interactions default to equilibrium be undermined. Second, the data from three of the major models used in RL point to the strong influence the underlying algorithms have on the decision-making of such agents. Not only does the type of RL model influence the degree to which the agent will cooperate with the human, but it also shapes the responses the humans will have during such encounters. Repeated interactions between humans and AIs further underscore the divergent behavior of each RL model when playing against humans. Third, the experiments yielded strong evidence in favor of the notion that humans are influenced by biases towards AI. Even though the primary dynamic of each game is the Nash Equilibrium (since it's the mutual best response emerging from what each player expects the other to do), the human players attribute significance to whether they are playing against an AI or not when they develop expectations of cooperative behavior.

It's thus important to note that AI safety researchers should not assume that the willingness for an AI to cooperate with humans in one scenario necessarily generalizes to all situations and vice versa. Our setup goes a long way in establishing a strong basis to investigate human-machine teamwork by testing such cooperative dynamics across a variety of games where coordination is in the collective interest of the multi-agent system. Using game theory in this setting is really useful because sharp deviations from Nash Equilibria are indicative of the complex nature of the interactions. However, limiting empirical research on human-machine cooperation to just one game would have only provided narrow evidence about the behavioral patterns of both AI and human players. Thus, testing different RL models, different beliefs about whether the other player is an AI or a human, as well as different game theory models with different Nash Equilibria provides a much stronger basis to make inferences about the ability for human-machine teams to coordinate.

## 3.7 Conclusion

RL is quickly becoming the dominant machine learning paradigm because of its generalizability. Thus, questions about the role incentive structures play in human-machine cooperation are now more relevant than ever. To that end, this study's methodology shows how game theory models are an effective interface to empirically study the degree to which humans and machine cooperate under different circumstances as well as the factors that influence such cooperation. The next chapter will extend this methodology to look how principles of human-human teamwork influence human-machine cooperation.

# Chapter 4

# Study 2: Human-Machine Teams as Multi-Agent Systems

This study expands upon the methodology used in Study 1 in that it uses a game theory scenario (albeit a different one with a focus on coordination as opposed to cooperation). It addresses RQ1 by looking at how different aspects of team perception (as elicited by a team effectiveness survey) affect the human participant's ability to effectively coordinate with different types of RL agents. The results informed the methodology we are deploying more comprehensively in Proposed Study 1.

Reported here are major excerpts of the journal article that was subsequently submitted and is now under review.

## 4.1 Introduction

Game theory is the study of the decision-making behavior of reward-maximizing agents in strategic situations [Von Neumann et al., 2007]. It integrates economics and math to provide a framework to encode incentive structures into matrices. This property is extremely useful because AIs, unlike humans, understand their environments through matrices. Furthermore, the fundamental theorem of game theory developed by John Nash guarantees that in every game there will be a Nash Equilibrium (NE), a set of strategies where each player will converge to as they mutually respond to each other [Nash et al., 1950].

Very often, however, the NE leads to an outcome that is sub-optimal or collectively harmful to the players, and yet is inescapable precisely because it emerges from each agent not having a strictly better alternative given what they expect the other players to do [Bab and Brafman, 2008]. In such dilemmas, only a coordinated and joint strategy by all the players will result in optimal rewards for the group [de Cote et al., 2006]. Cooperation, fairness, reciprocity, and coordination and reciprocity rely upon the players' preferences, which are in turn systematically influenced by the game's incentives, and can be designed to tie rewards to collaborative outcomes [Erev and Roth, 1998].

The class of machine learning models that best responds to incentives and rewards is reinforcement learning (RL) [Tuyls and Weiss, 2012]. These algorithms employ behavioral models that reward and punish behavior to induce the discovery of a decision framework that maximizes positive rewards over time [Tuyls and Weiss, 2012]. The challenge RL agents face is that exploring new actions that can yield better values over time comes at the expense of committing to selecting the best action available at the time to maximize rewards [Tuyls and Weiss, 2012]. Regardless of such tradeoffs however, RL has been a driving force in major advances in AI in recent years, in a variety of settings, such as Go, Chess, soccer, and Atari games [Hu et al., 1998, Foerster et al., 2018].

However, a major challenge to RL emerges when the agents are placed in an environment where multiple NEs exist, since learning one NE strategy does not necessarily result in an optimal outcome because there is no guarantee the other players will also select the same NE out of the possible ones [Hu et al., 1998]. Very often, the agents in multiplayer settings remain trapped in game-theory-induced dilemmas. Furthermore, RL agents over-estimate rewards in zero-sum games and fail to learn to reliably cooperative over time by not considering the other player's learning process [Tampuu et al., 2017, Foerster et al., 2018]. Thus, game theory provides an effective interface to study the cooperative dynamics and the influence of incentives in multi-agent systems of humans and AIs.

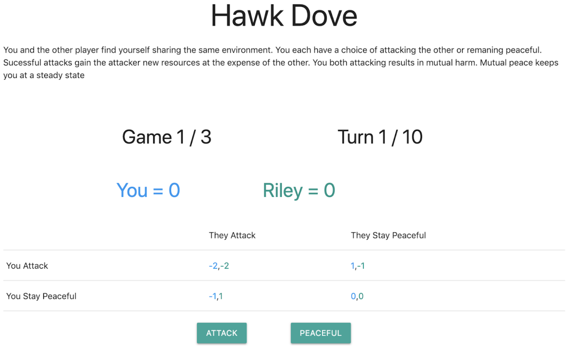
The next aspect to consider is whether teams that include AI teams can collaborate as effectively as human teams do. Specifically, whether AIs are capable of larger scale cooperation is still an open question. Prior research has shown that bots attempting to solve a graph coloring problem (a high-coordination-required matching task) successfully coordinate on a micro-level but fail to do so on a macro-level, thereby preventing the collective goal of a stable match to occur [Shirado and Christakis, 2017]. However, the study also demonstrated that adding humans to the team accelerating the median solution time by 55.6% [Shirado and Christakis, 2017]. Furthermore, integrating human factors in the design of human-machine teams has led to remarkable results in the healthcare domain. Specifically, teams of oncologists partnered up with machine learning algorithms outperformed both state-of-the-art neural networks as well as expert teams in diagnosing cancer [Wang et al., 2016]. These results speak to the major advantages that emerge from harvesting and integrating human intelligence and AI, and should pave the way for more research of this kind.

## 4.2 Hawk-Dove Games and Zero-Sum Mentalities

The Hawk-Dove game (HD) is a game where each player is faced with a decision of whether to attack or to remain peaceful. Each player operates under the same incentive structure, where collective peace results in both players receiving 0 points and collective war resulting in both players losing 2 points. The key aspect of the game is that each player is better off being peaceful, unless they manage to successfully attack when the other player goes for peace.

Figure 4.1 shows the information available to the human players when they are assigned to play the HD. A successful attack results in a point being transferred from the peaceful player to the aggressive player. It's important to note that such an outcome leads to a smaller lose for the peaceful player than in a war (mutual aggression), which in turn creates a powerful incentive to remain peaceful unless either of the player is driven by a zero-sum mentality.

Figure 4.1: Hawk-Dove Interface



Hawk Dove

You and the other player find yourself sharing the same environment. You each have a choice of attacking the other or remaning peaceful. Sucessful attacks gain the attacker new resources at the expense of the other. You both attacking results in mutual harm. Mutual peace keeps you at a steady state

Game 1 / 3 Turn 1 / 10

You = 0 Riley = 0

| | They Attack | They Stay Peaceful |
|---|---|---|
| You Attack | -2,-2 | 1,-1 |
| You Stay Peaceful | -1,1 | 0,0 |

## 4.3 Experimental Setup

Experimental design details include enlisting over 100 participants from Amazon Mechanical Turk to each play the HD with RL agents. We also collected demographics and a survey of team effectiveness and perceptions at the end of every game, as well as the outcome data from each game from the web interface.

| | Male | Female |
|---|---|---|
| Gender | 23.77% | 76.23% |

| | African American | Asian | Caucasian | Other |
|---|---|---|---|---|
| Race & Ethnicity | 3.52% | 54.64% | 40.08% | 1.76% |

| | 18-25 | 26-35 | 36-45 | 46-55 | 56-65 |
|---|---|---|---|---|---|
| Age | 25.55% | 53.73% | 10.13% | 6.16% | 4.40% |

| | Doctorate (PhD, EdD) | Professional Degree (MD, DDS, DVM) | Associates Degree (AA) | High School (GED) |
|---|---|---|---|---|
| Education | 0.44 | 4.40 | 5.28 | 7.04 |

| | Some College | Bachelor's Degree (BA, BS) | Masters (MA, MS, MEd) |
|---|---|---|---|
| Education | 7.04 | 62.99 | 12.77 |

Table 4.1: Demographics

Figure 4.1 displays the interface we built for the experiment. It records and stores each move made by each player as they play 3 rounds of 10 turns. Our game interface automatically collected userID, the RL agent's move (0 for cooperation, 1 for defection), the human player's move (0 for cooperation, 1 for defection), the type of RL agent (”ppo”, ”vpg”, ”dqn”), and the turn number.

We used the TensorForce open-source library to implement the RL agents in a modular way [Schaarschmidt et al., 2017, Schaarschmidt et al., 2018]. We focused on Vanilla Policy Gradient (VPG), Deep-Q Networks (DQN), and Proximal Policy Optimizers (PPO) for our experiment and results [Williams, 1992]. DeepQ agents rely on Q-learning and yet overcome its limitations through the use of deep neural networks to estimate value for unseen states [Schulman et al., 2017]. Once the experiment was completed, each pair of moves was translated into an outcome (cooperation, non-cooperation), and each move by the human, as well as the RL agent, was also classified as cooperative or non-cooperative. The survey results were then matched to the participant's ID.

The survey enabled participants to answer Likert-scale questions in regard to their feelings towards AI and their perception of different team aspects of the task. The analysis focused on a subset of survey data in light of the length restrictions. The data being analyzed focused on the following questions:

1. ”I felt my teammate and I had a shared understanding of our teamwork”
2. ”I am optimistic towards AI”
3. ”I would prefer working with a human rather than an AI”

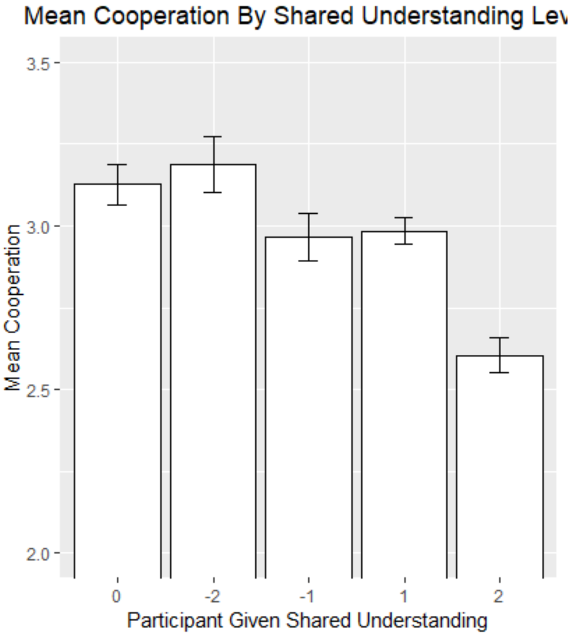
## 4.4 Results of the Game Outcomes and Survey Data

Cooperation was associated with participants' level of agreement towards three different phrases using Likert scales. The first phrase, “I felt my teammate and I had a shared understanding of our teamwork,” was used to see if there was a relationship between a participant's perceived level of shared understanding, and their actual level of cooperation with their teammate. The second phrase, “I am optimistic towards artificial intelligence,” was used to see if a participant's feelings towards artificial intelligence would be correlated with their level of cooperation with their teammate. The third and final phrase, “I would prefer working with a human rather than an AI,” was used to see if a user's feelings towards an AI teammate could hinder their overall cooperation. All of these scales were used with levels of cooperation to create ordered logistic models.

### 4.4.1 Shared Understanding

Participants rated their perceived level of shared understanding with their teammate by responding to the phrase “I felt my teammate and I had a shared understanding of our teamwork” on a scale from very little (-2) to very much (2). Using neutral (0) as a baseline, ordered logistic regression models were created to determine the relationship between a user’s perceived level of shared understanding and the actual team cooperation they experienced.

Figure 4.2: Mean cooperation based on different levels of shared understanding



| Shared Understanding (0 baseline) | CI: 2.5% | 97.5% | Odds Coefficient |
|---|---|---|---|
| -2 | 0.9105888 | 1.3602221 | 1.1123634 |
| -1 | 0.6542316 | 0.9199658 | 0.7758380 |
| 1 | 0.7136371 | 0.9368801 | 0.8179720 |
| 2 | 0.3652745 | 0.4896615 | 0.4230801 |

Table 4.2: The relationship between perceived shared understanding and cooperation

This model was able to determine a significant relationship between shared understanding and a team's level of cooperation

$$(\chi^2 = 216.866, p << 0.001)$$

. Users who gave shared understanding a -1, 1 or 2 were associated with a significant (Confidence Interval (CI) does not cross 1) difference in levels of cooperation than those with a neutral response towards shared understanding. From this model, the level of a team's cooperation sees a strong association with the level of shared understanding given by a participant. The only levels of shared understanding are associated with a decrease in the odds of increasing team cooperation (Table 4.2). This level of shared understanding may be affected by other factors, like optimism and human preference, but it is still shown to be an important factor in increasing cooperation. However, the best level of shared understanding, supported by the data, is a neutral understanding; this could be due to feelings of overconfidence or underconfidence in the level of shared understanding between human and AI teammates.

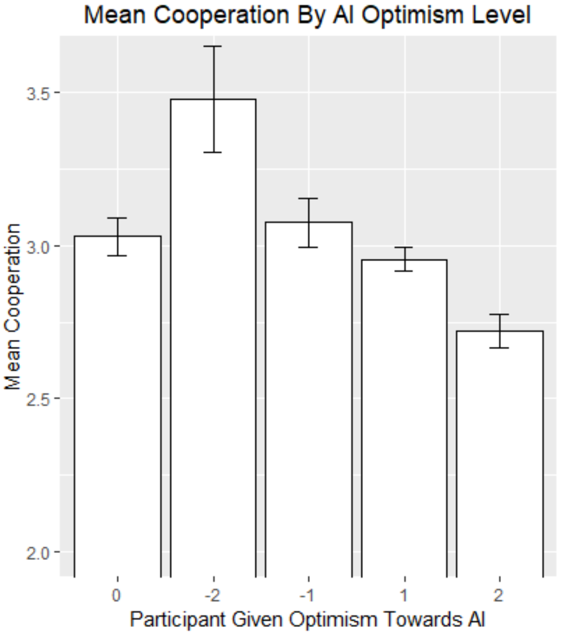
### 4.4.2 Optimism Towards Artificial Intelligence

Similar to shared understanding, participants were asked their level of optimism towards artificial intelligence by responding to the phrase "I am optimistic towards AI" on a scale from very low (-2) to very high (2).

| Shared Understanding<br>(0 baseline) | CI: 2.5% | 97.5% | Odds Coefficient |
|---|---|---|---|
| -2 | 1.5371439 | 3.6239674 | 2.2336973 |
| -1 | 0.9544499 | 1.3767417 | 1.1460139 |
| 1 | 0.8247815 | 1.0695927 | 0.9394992 |
| 2 | 0.5440443 | 0.7277647 | 0.6293811 |

Table 4.3: The relationship between optimism towards AI and cooperation

Figure 4.3: Mean cooperation based on different levels of optimism towards artificial intelligence



The model saw a significant relationship between the participant's optimism towards artificial intelligence and their overall team cooperation

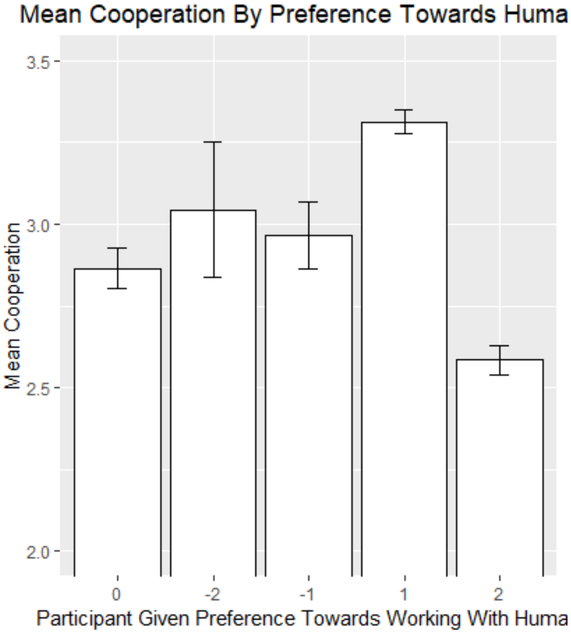
$$(\chi^2 = 2760.496, p << 0.001)$$

. Users who had levels of optimism -1 and 2 saw a significant (CI does not cross 1) difference in overall team cooperation versus those with a neutral level (Table 4.3). Significant differences in team cooperation can only be observed at the extreme ends of this scale. This leads to the conclusion that Optimism may only affect team cooperation at very strong levels. The most cooperative teammates were seen to have the lowest levels of optimism towards AI, this could be due to the human working harder to cooperate due the lack of trust in the AI, while humans with high optimism may put too much trust into the AI lowering the overall cooperation.

### 4.4.3 Preference Towards Human Teammates

Like the previous two phrases, participants rated their opinion towards the phrase “I would prefer working with a human rather than an AI” on a scale from very little (-2) to very much (2). These responses were used to see if a preference towards a human teammate could possibly be a contributor towards overall team cooperation.

Figure 4.4: Mean cooperation based on different levels of preference towards human



| Shared Understanding<br>(0 baseline) | CI: 2.5% | 97.5% | Odds Coefficient |
|---|---|---|---|
| -2 | 0.9050088 | 2.0107818 | 1.3418218 |
| -1 | 0.9670076 | 0.4552385 | 1.1856354 |
| 1 | 0.9865519 | 2.5747484 | 2.2615868 |
| 2 | 0.5583562 | 0.7159055 | 0.6323584 |

Table 4.4: the relationship between preference towards working with a human and cooperation

The model saw a significant relationship between a participant's preference towards a human teammate and their levels of cooperation

$$(\chi^2 = 3259.792, p << 0.001)$$

. Users who had a preference level of 1 or 2 saw a significant (CI does not cross 1) difference in overall cooperation versus those with a neutral level (Table 4.4). Although participants with a slight preference towards human teammates saw an increase in the odds of having higher cooperation, participants with a strong preference towards human teammates actually saw a decrease in cooperation. This observation leads us to believe that strong preferences towards human teammates could be associated with decreased cooperation, while a slight preference towards human teammates could yield human teammates willing to work harder to cooperate to make up for faults they see in AI teammates.

## 4.5 Discussion

The results suggest that human participant's preferences towards AI have a substantial effect on whether or not the human player decides to be cooperative. Cooperation occurs more frequently when participants don't hold strong views over whether they'd rather play with a human or an AI, which suggests that biases against AI are not necessarily wide-spread but do affect the human willingness to cooperate in the hawk-dove scenario.

The other major result is that there is an extremely significant relationship between perceived shared understanding and cooperation. Such results validate one of the major components of human-human teamwork, as it's shown to also apply to human-machine teams. These results speak to a different dimension of human-machine cooperation than personal biases towards AI. Shared understanding is an emergent property of teamwork, thus our results can pave the way to study other factors involved in successful human-human teamwork [McNeese et al., 2017]. We should not assume that all the principles of human teamwork would be validated in the human-machine setting given the limited communication involved, yet our results with shared understanding suggest that some aspects would overlap between the two types of teams.

The underlying dynamic of the hawk-dove game should also be underscored: both players are better off remaining peaceful but can only do so if they are willing to forego the short-term incentive to betray each other [Neugebauer et al., 2008]. Prolonged periods of cooperation that we have observed in our data are in some ways irrational, for the expectation of peaceful behavior by the other player would induce a self-interested agent to attack. The fact that this is not the case speaks to the ability for humans and AIs to coordinate through their behavior even when communication is not possible, such as in our setup.

The cooperative behavior that leads to coordination and mutually beneficial outcomes is thus mediated by many of the factors that lie at the core of team effectiveness, such as shared understanding. For future work, a more expansive survey could be delivered to infer a broader spectrum of characteristics and human qualities that are conducive to human-machine teamwork. The methodology deployed in the study can be a very useful test-bed to empirically verify principles of teamwork in human-machine interactions because it leverages game theory's role in both human as well as AI decision-making.

## 4.6 Conclusion

Understanding human-AI teams requires bridging the gap between human teamwork and multi-agent systems. To that end, this study paired human participants with reinforcement learning agents to play a game theory scenario that emphasizes cooperation. Results indicate that human biases in favor or against AI have significant effects on aggression and peacefulness, and that shared understanding plays a major role in cooperation. The next chapter shifts the focus towards theory by introducing the first theoretical method we developed that will be tested in this dissertation.

# Chapter 5

# Model A: Human-Centered Prediction Markets

This chapter outlines the theoretical foundation of the model we are implementing in Study 4. By approaching prediction markets from a human factors perspective, it provides a useful model to address RQ4, RQ5, and RQ6 by identifying the ways in which AI can play a role in prediction markets and its relevance to collective intelligence. Moreover, it provides a framework to interpret the results we are anticipating from Proposed Study 2 and grounding them in a human-machine perspective through the literature on collective intelligence.

Much of the work contained in this chapter refers to and is taken from what became the final paper that was eventually published.

## 5.1 Introduction

Prediction markets are mechanisms that enable participants to bet upon the occurrence of particular events [Hanson, 2003]. The objective of a prediction market is to create an incentive structure to coordinate a sophisticated forecasting process that can enable organizations, communities, and countries to better deal with uncertainty about the future. To that end, they have been studied in many social psychology contexts to identify their connections to collective intelligence [Tetlock and Gardner, 2016].

However, despite their relevance to cognitive science, prediction markets have been mostly analyzed in the context of economics and computer science, where the objectives are optimizations of the underlying process as opposed to collective intelligence.

We approach prediction markets from a human factors perspective to identify the key cognitive features that enables collective intelligence to emerge. Subsequently, we propose a new human-factors-based prediction market whose design enhances collectively intelligence. Furthermore, we go one step further and integrate an artificial intelligence component to set the foundation for much higher degrees of collective intelligence. Thus, our model enables designers to leverage all the recent advancements in machine learning.

## 5.2 Model Design

Our prediction market in Figure 5.1 relies on 4 components: 1) *human participants* without access to historical trading data, 2) *noisy trading bots* that trade randomly to generate a profit opportunity for other traders, 3) *market making bots* that learn from market patterns and trade in order to stabilize or open up new markets for event, and 4) *neural network* that receives the data generated by the prediction market to make its estimate.

As Figure 5.1 shows, the noisy bots create the profit opportunity that motivates the human participants to engage in the trading alongside the market making

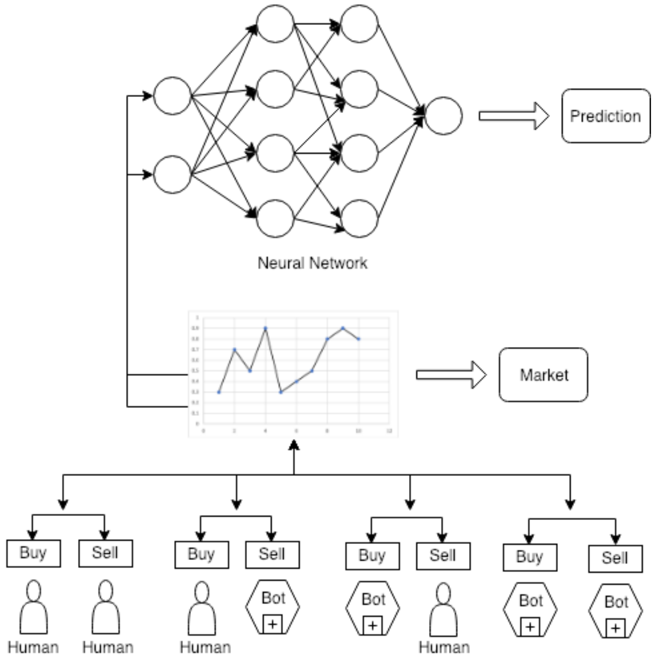


Figure 5.1: The collective intelligence emerging from humans and AIs in the prediction market becomes the input of a neural network

bots that rapidly smooth price changes and stabilize the market for each event. The data generated by the process is then fed as the input to the neural network that calibrates its aggregate estimate of all the participants over time.

### 5.2.1 Human Interface

The primary objective on the human side of prediction markets is to incentivize only traders with new information to engage. Without such a goal, the market's incentives can lead to speculate behavior observable in stock markets, where some traders specialize in trading on price movements as opposed to changes in fundamentals. Such behavior would dilute the signal of prices in the prediction market for they would be distorted by copycat and speculative traders that are not contributing to the implicit deliberation process. To that end, our design does not include interfaces displaying historical trading data as to induce traders to consider only information relevant to the reality of the event as opposed to its financial counterpart in the prediction market.

Furthermore, we deploy Hanson's market scoring rule market maker (MSR) to generate activity on new events [Hanson, 2003, Chen et al., 2010]. This structure induces even just a single trader to reveal their information, which would otherwise not occur under a standard double action in traditional prediction markets. Furthermore, MSR enables the manager of the prediction market to not only reduce the amount of money needed to simulate initial activity in the prediction market, but to also efficiently allocate that capital in fixed amounts set in advance regardless of how active the trading ends up being before the final estimate [Chen and Vaughan, 2010].

From a human factors perspective, our design seeks to use these incentive structures to guarantee non-competitive self-selecting mechanism that encourages the type of diversity in information and decision-making needed for prediction markets to avoid the kind of exuberance that hallmarks "bubbles" in financial markets. In our prediction market, each participant is asked to evaluate the uniqueness of their information before entering into any trade. This is further reinforced by the reasoning from the current market price as to whether it reflects the participant's unique information or whether it justifies the buying or selling of shares to direct the price closer to their estimate. Diversity has value in a prediction market, thus participants with diverse information will self-select to become a trader [Watkins, 2007].

Beyond that however, our design incorporates differences in forecaster knowledge and skill. Specifically, the order size and the amount of money at stake on a given trade serves as a useful proxy for the participant's confidence, for their risk-aversion will make the amount of invested be proportional to the gap between their expectations and the current market price. On a long enough time line, successful traders will be highly rewarded and thus wield greater influence on future prices. This feature is consistent with the “marginal trader hypothesis” in economics where the efficiency of a stock market is driven by a minority of unbiased and active participants who wield corrective influence [Wolfers, 2009].

### 5.2.2 Machine Interface

The machine side of the design includes two types of agents: noisy traders and market makers. We designed them because they provide distinct functions within the prediction market, by making it more effective and precise.

The noisy traders are designed to buy and sell shares of each event randomly and selling them within a narrow interval from their current price. Their behavior creates the opportunity for the human participants to profit on a regular basis, preventing a no-trade scenario that would make the price updates given new information sudden and extreme as opposed to gradual and incremental [Chen and Vaughan, 2010]. This type of agent resolves many of the concerns related to motivation of the human participants.

The market makers are altogether different. They are not designed to behave randomly, but rather learn how to trade from the historical trading data of the prediction itself by using simple machine learning algorithms. Their purpose is to stabilize trading by engaging in the type of behavior high frequency trading firms in financial markets engage in to drive out speculation on price movements. Their function is to also efficiently make bets on forecasts two-sided to reduce imbalances, and the nature of machine learning positions them to rapidly improve their usefulness as they receive new data from the human participants over time. Their trading behavior is extremely important to counterbalance the long-shot biases that plague human-only prediction markets.

Furthermore, prior research with Football forecasts indicates that even though such bots have no understanding of the underlying event being analyzed, they are on average more accurate than the human participants and thus force the participants to refine their analysis which improves their forecasts [Malone, 2018]. This type of research suggests that prediction markets would benefit from the type of machine interface we include in our design to mitigate human biases and create a more robust guarantee that prediction markets focus on forecasting as opposed to speculation.

### 5.2.3 Neural Network Layer

A major contribution of this model is the layer that sits above the human participants and the trading bots. We include a neural network whose input is the data generated by participants to produce an aggregate estimate of the probability of the event occurring. Artificial neural networks, much like the human brain, use neurons that are made up of collections of nodes that function as processing units with weighted connections to each other [Kaur and Wasan, 2006]. A neural network has a very basic architecture: it has an input layer of neurons that accepts input, a customizable number of hidden layers performing calculations and transformations over the data, and an output layer of neurons that outputs predictions in the selected format [Kaur and Wasan, 2006]. Whenever a neural network makes a prediction, the error rate is measured so that the network can adjust the weights of its neurons in order to calibrate its model and achieve better accuracy over time [Simon and Eswaran, 1997].

As mentioned earlier, there is no real consensus on how to translate fluctuating prices in the prediction market into a sensible probability estimate for the event being forecasted. We believe the issue lies in the prior literature emphasis on algorithmic solutions. Instead, we seek to integrate the recent advances in AI to create dynamic as opposed to static solution to the problem, and one that evolves with the prediction market over time.

Prior research has been able to show at a theoretical level how the market scoring rules of prediction markets connect to the equations underlying the value of function of many machine learning models (Chen  Vaughan, 2010). Beyond the theory however, artificial prediction markets can fuse the predictions of trained classifiers into contract prices on all possible outcomes [Barbu and Lay, 2012]. The results indicate that such systems outperform cutting edge algorithms that ensemble a variety of models in the healthcare domain, which is attributed to the market mechanism's ability to aggregate specialized classifiers that participate only on specific instances [Barbu and Lay, 2012, Jahedpari et al., 2017].

This type of result however extends beyond artificial prediction markets where machine learning models are the participants. Tetlock's research on forecasting for IARPA also showed that an extremizing algorithm that took the probabilistic estimates of "superforecasters" as its input actually outperformed 99% of the individual super-forecasters, by aggregating their opinions and weighing them based on track record and diverse POV [Tetlock and Gardner, 2016].

We go multiple steps further with our design. Not only do we accommodate a hybrid prediction market where humans and AIs alike participate, but we also reject static algorithms in favor of a neural network which is much more effective and capable of aggregating and weighing the different estimates and viewpoints emerging from the interactions of the prediction markets. The neural network will not merely be learning from price fluctuations, but from the trading behavior of the human participants as well, potentially identifying talented forecasters without having to rely or give undue influence to any of them. Furthermore, our design generalizes to different settings because the neural network obviates the need to perfect knowledge from the group as it detects patterns among participants who themselves may not be experts.

In essence, the machine learning layer enables our prediction market to transcend the limitations inherent to traditional prediction markets by having a neural network train and learn from the data generated by the human traders.

## 5.3 Discussion

Prediction markets have been proven useful in forecasting geopolitical events, sports outcomes, and elections. Yet, they have failed to become ubiquitous despite their success. We believe this is due to the many flaws of traditional prediction markets that privilege algorithms over human effort. By putting collective intelligence at the center of the discussion, our design shows us how a human factors perspective can not only enhance human intelligence, but also make room for artificial intelligence.

On multiple fronts, AI can be the key to overcoming many of the challenges posed by prediction markets. It can downplay the influence of human biases in the market by checking and balancing activity to suppress bubbles before they form. It can induce the participants to think probabilistically and more precisely about their understanding of the event. It can also make the market more adaptive to new information by forcing every participant to update their beliefs based on new information. Through these interventions, the machines in our prediction market smooth out the often erratic behavior of human traders and thus provide a more reliable forecasting mechanism that can extend beyond the few areas where prediction markets have been tried.

Beyond that, however, our model forces us to reconsider how we think about prediction markets in the first place. By de-emphasizing the computation and game theoretical perspective in favor of the cognitive science one, we no longer think about prediction markets as an auction mechanism but rather as a platform for decentralized collaboration between humans trying to tackle the challenges of uncertainty.

Furthermore, our design opens up a new branch of research where the prediction market is seen as a coordination mechanism to enable a different type of cognition: artificial intelligence emerging from collective intelligence. The wisdom of the crowd becomes the input the neural network learns from, whose emergent property is an altogether different type of artificial intelligence that is worth exploring in future research.

## 5.4 Conclusion

Through an interdisciplinary perspective, we brought a human factors approach to the design of a more effective prediction market that is not merely optimized for the wisdom of the crowd but also enables a higher level cognitive process that integrates collective intelligence with artificial intelligence. The next chapter introduces another model we developed for human-machine teamwork that will be tested through this dissertation.

# Chapter 6

# Model B: Collectively Intelligent Teamwork

This chapter outlines the theoretical foundations of one of the major underlying theses of the dissertation: team cognition and collective intelligence are analogous phenomena happening at different scales. It classifies the similarities and differences in the cognitive models represented in Study 3 (team cognition) and Study 4 (collective intelligence) in order to identify the connections between them to set the stage for them to be combined into a more unified understanding of emergent cognition. Furthermore, this chapter also discusses the Superforecasting phenomenon and its several major implications for human-machine intelligence.

Much of the work contained in this chapter refers to and is taken from what became the final paper that was eventually published.

## 6.1 Introduction

Cognitive science is a study of information processing and neurological behavior that all take place in the human mind. Through those lenses, it becomes possible not only to examine individual behavior, but also to investigate social behavior. Team science lies at the intersection between researching individual and group psychology, and actively deals with how collections of individual people coordinate most effectively.

This kind of phenomenon however is replicated at a larger scale with intellectually productive crowd efforts. A primary example is Foldit, a crowdsourcing effort for biochemistry and protein folding that uncovered in just three weeks the structure of an enzyme related to AIDS that had eluded scientists for 15 years [Malone, 2018]. This type of phenomenon taps into resources and skills needed to perform an activity that are distributed widely or reside in places that are not known in advance [Malone et al., 2009].

So far, the team cognition literature has mostly focused on the ability for teams to develop shared situational awareness or a shared understanding of a complex task in many practical contexts [Cooke et al., 2007]. On the other hand, the collective intelligence literature has mostly focused on the ability of large decentralized groups to aggregate and process large amounts of information to produce insights far superior to those of individuals and teams, especially in scenarios marked by deep uncertainty [Atanasov et al., 2016].

In this model, we analyze the results of "superforecasting" (teams of forecasters outperforming prediction markets) as the starting point for a new way to think about teamwork that merges the insights from team cognition and collective intelligence to challenge many assumptions held by each school of thought. Thus, one of the contributions of this model is a comparative analysis of the team cognition and collective intelligence literature to identify the connections between the concept of a shared mental model and that of the wisdom of the crowd respectively.

Furthermore, we analyze the results from the superforecasting research to develop a model that enables the coordination individuals whose information processing enables both team cognition as well as collective intelligence to emerge. Lastly, we demonstrate how to incorporate artificial intelligence into our model in order to enable team's information processing to become the input of a neural network to enhance the team's collective intelligence to outperform smart crowds.

## 6.2 Model Description

On the surface, team cognition and collective intelligence appear to be irreconcilable. Teams cognition is heavily dependent on information sharing and verbal communication, whereas collective intelligence relies on the opposite: local information gathering and independent analysis [Graham et al., 2004]. Team cognition benefits from centralized leadership, whereas collective intelligence is predicated upon decentralization [Fiore et al., 2010]. Teams exaggerate the cognitive biases of individuals, whereas wise crowds compensate for them [Atanasov et al., 2016].

However, by delving deeply into the phenomenon of "superforecasting", a new type of team cognition can be identified as the properties of the wisdom of the crowd can emerge from small teams operating in the right structure. In turn, this insight sets the foundation for methods to combine artificial intelligence with collective intelligence. We thus develop a new model (Figure 6.1) for teamwork that unifies team cognition, collective intelligence, and artificial intelligence.

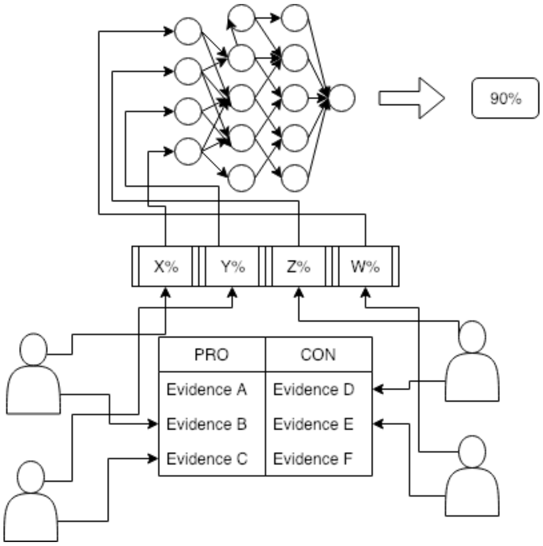


Figure 6.1: Each teammate shares evidence for a team discussion, and expresses a probability estimate that is then aggregated by the neural network

## 6.3 Collectively Intelligent Teams

A collectively intelligent team is a team whose shared mental model enables higher degrees of collective intelligence than in crowds. The first major component is a set of cognitive strategies that are deployed at the individual level. Specifically, each individual is trained on specific cognitive strategies to improve their decision-making. Just like superforecasters, collectively intelligent teammates are trained in probabilistic thinking, evidence-based discussions, depersonalized intellectual conflict, diversity of viewpoints, commitment to truth, and data-driven decision-making. All of these practices can dramatically enhance the teammate's team cognition.

The second major component is a set of guidelines to the team for information processing. Elite superforecasters in the literature are shown to engage in many distinct behaviors that were predictors of accuracy, they cognitively triaged (deciding how to allocate effort across questions). They asked five times more specific questions than average and were answered six times more frequently. They made comments roughly one third longer, and made between nine and thirteen times more general comments than average. Beyond merely being more likely to gather news and opinion pieces related to the forecast, superforecasters were between six and ten times more likely to share news links with their teammates [Mellers et al., 2015].

The last component is a layer of artificial intelligence to aggregate the team's insights. An often overlooked result of the superforecasting research is how applying an extremizing algorithm that aggregates the forecasters' forecasts and weighs them based on track record and diverse point of view outperformed 99% of the individual super-forecasters [Tetlock and Gardner, 2016]. Already prediction-polled superforecasters outperform prediction markets, the best collectively intelligent mechanism to date, so the fact that shifting from a statistical aggregation rule to an algorithm leads to even better predictive power is remarkable. Prior research in both prediction markets and prediction polling has identified the choice of an aggregation function as a factor that materially changes the value of the probability estimates of the system [Atanasov et al., 2013, Atanasov et al., 2016].

The advances brought about by a simple algorithm open up the opportunity to incorporate AI in the forecasting process. Prior research has shown how collectively intelligent teams can be created through cognitively optimized software applications. Specifically, a recent radiology study at Stanford has shown that a team radiologist coordinated with a probabilistic interface gained a 22% margin on the state-of-the-art deep-learning solution and a 33% margin on individual radiologists as a whole [Rosenberg et al., 2018b, Rosenberg et al., 2018a]. This result is not isolated to healthcare but also works in financial forecasting. A recent Oxford study on using the same probabilistic interface to coordinate traders to predict four economic indicators showed a 26% increase in prediction accuracy as compared to individual forecasts [Rosenberg et al., 2018b, Rosenberg et al., 2018a].

However these results occur with the use of simple algorithms that aggregate the individual forecasts to produce an optimal estimate. Neural networks can process much deeper correlations between individual forecasts over time, therefore machine learning can yield significant gains for collectively intelligent teams that are orders of magnitude better than base performance.

In essence, our model for collectively intelligent teamwork relies upon each teammate being trained by probabilistic cognitive strategies, the team sharing information through protocols that preserve independence, and most importantly the use of neural networks to optimize the collective intelligence of the team.

## 6.4 Discussion

The superforecasting research could not be understated: those results robustly demonstrate how small teams can outperform the most collectively intelligent mechanism known to date – prediction markets – in a significant way. This evidence suggests that a new type of intelligence can be unlocked through teamwork that goes beyond what the team cognition literature has been demonstrating. Team cognition expands from perception into prediction as a team of superforecasters displays both team cognition and collective intelligence, two emergent phenomena thought to be separate up until that point.

Researching superforecasters expands how we think about collective intelligence as well as team cognition. For the former, it means that you no longer necessarily need a crowd to have collective intelligence as long as teams are trained through particular cognitive strategies when engaging in forecasting. For team cognition, it means that the possibility landscape for teamwork is vaster than previously thought, and is relevant beyond the military setting and responding to physical situations and can move into a higher-order form of information processing like forecasting.

Furthermore, our model bridges the literature gap between team cognition and collective intelligence. Specifically, we note that shared mental models are to team cognition, what the wisdom of the crowd is to collective intelligence: they are emergent phenomena of effectively coordinated groups of people. Through the collectively intelligent team model, these two perspectives can be integrated. Not only can teams rapidly develop a sophisticated understanding of their situation and environment, but can also work together to produce remarkable insights under deep uncertainty by thinking about the future in a probabilistic manner.

The collectively intelligent team model is also useful as the foundation for new technology. A collectively intelligent team emerges when a team's shared mental model enables each teammate to enhance their forecasting process so that they can each produce better estimates that can be fed as input to a neural network that learns over time how to calibrate the weight it assigns to each team member's opinion in particular contexts. The use of AI in this case creates a highly adaptive coordination mechanism that enables the team to retain independent thought while still collaborating. Rosenberg et al's (2018) work on coordinating teams through probabilistic decision making is merely an initial attempt at developing human-centered interfaces that can transform team cognition into collective intelligence.

Interestingly enough, the literature on team cognition provides many results that are consistent with collectively intelligent teamwork. For example, Google's Project Aristotle illustrates how the stronger predictors of team success in knowledge-work shows are equality of conversational turn-taking and higher levels of social sensitivity [Duhigg, 2016]. This is consistent with the observations that superforecasters deeply value hearing everyone else's opinion before expressing their own [Tetlock and Gardner, 2016].

Furthermore, prior research has shown that asking team members to defend their position induces a cognitive strategy relying on generative confirmatory evidence, whereas not expecting them to leads to greater freedom to explore arguments counter to initial positions [Hinsz et al., 1997]. Teams of superforecasters do this by keeping track of arguments from different viewpoints in favor or against a particular forecast [Tetlock and Gardner, 2016].

## 6.5 Conclusion

Through a deep analysis of team cognition and collective intelligence, we developed a new model that integrates the two phenomena. This model sets the basis for a new type of teamwork that exhibits both properties, as evidenced by the results on teams of superforecasters. Lastly however, we go one step further to show how artificial intelligence can be used to cognitively enhance collectively intelligent teams to reach a new level of intelligence that has only begun to be fully explored.

# Chapter 7

# Study 3: Team Cognition in Human-Machine Teams

## 7.1 Introduction

There are many ways of studying teamwork, but only a few that can identify the properties of team cognition. This study focuses on the first research goal: understanding team cognition in human-machine teams. Team cognition is an emergent phenomenon of shared understanding that arises out of the coordination of the team members as they try to adapt to a complex environment and deliver on a challenging task. In order to study teamwork and team cognition, it's thus not only useful but necessary to create a dynamic environment that motivates the teammates to work together in order to accomplish a goal. To that end, simulations and simulated environments can effectively provide such a set up in a way that is replicable across iterations, so that team performance can be meaningfully compared.

## 7.2 Overview

The study analyzes three major types of teams (human teams, machine teams, and human-machine teams) from both a quantitative perspective (looking at the comparative performance of each team type) and descriptive perspective (using teamwork surveys to identify principles of human teamwork that are predictive of human-machine team performance). This dual approach addresses RQ2 and RQ3 with the quantitative analysis, and RQ1 with the qualitative analysis.

This between-subjects study uses an emergency response simulation to create an experimental environment where participants have to work together to coordinate effective team responses to a variety of situations. Traditionally, the emphasis has been placed upon the role communication plays into the development of a shared mental model [Graham et al., 2004, Mathieu et al., 2000]. However, the type of nuanced and deep communication that is associated with effective human teams is not currently applicable in the case of human-machine teamwork. Thus, the emphasis in our study is not on communication, but rather on shared situational awareness (defined in our measurements section), as the humans and AIs have to work together by effectively anticipating each other's behavior and succeed as a team.

## 7.3 Methods

### 7.3.1 Simulation

The simulation we are using for this study is NeoCITIES. In NeoCITIES, participants respond to a sequence of incidents in a fictional college town that requires emergency response [McNeese et al., 2014a]. Teams of three individuals playing separate roles (police, fire, and hazardous material response) are tasked with assessing each episode and appropriately allocating scarce resources. This basic setup has historically lead to multiple findings in information science, psychology, and geographical sciences [McNeese et al., 2014a]. Furthermore, through the repeated use of established measures and metrics, NeoCITIES enables researchers to investigate specific aspects of team cognition and other critical factors in teamwork [Mohammed et al., 2010].

### 7.3.2 Updated NeoCITIES

In order to conduct our study, it became necessary to develop a new version of NeoCITIES that could incorporate AI agents and support a wider variety of scenarios. To better understand the nature of the changes that we made in the new version, it's important to begin by discussing the limitations of the prior version of NeoCITIES.

#### 7.3.2.1 Backend and Architecture

With respect to its architecture, the prior version of NeoCITIES used a client-server model to display information generated by a simulation engine that was developed with Java and Adobe technologies as a multi-threaded application that handles the event and resource data, maintains the state of the world, and calculates event scores [Hellar and McNeese, 2010].

The downside of this type of design is that it did not support multiple simulations simultaneously, and thus limited researchers to running one simulation at a time on their local environment. To solve this problem, we implemented a cloud-based design where clients push updates to a cloud server (specifically Firebase) that automatically updates a global state that the clients are listening to through a REST API. Essentially, any time changes are made to the server, the state of the simulation is updated and the clients reflect those updates in real time.

Furthermore, the Firebase server supports a vast amount of simultaneous sessions because each session is managed by a child in the database's JSON tree, which prevents the clients from accidentally updating the wrong session. Therefore, the new architecture leverages the benefits of the cloud to enable researchers to conduct simulations without having the participants be co-located, while also removing limitations on session being run simultaneously, thereby accelerating the data gathering process for researchers in the future.

The last major way in which the new architecture advances NeoCITIES is by creating an API layer that enables AI agents to easily interact with the human participants, thereby enabling the platform to support human-machine teams for the first time. Specifically, the AI agents can request the state of the simulation through the REST API on Firebase and push actions to the database. This enables the agents to participate in the simulations in an asynchronous manner similar to human players.

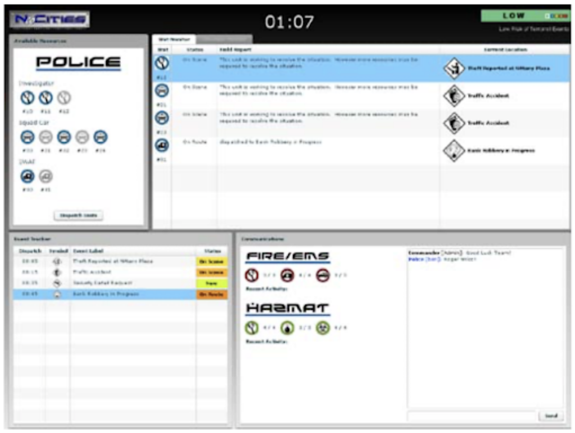
Overall, the new architecture is designed to be very modular and thus support researchers to develop a variety of scenarios with different roles, resources, and number of players. Prior versions of NeoCITIES required updating the backend every time a new scenario had to be implemented, thereby rendering all the parameters static. The new version sidesteps the entire issue by making the scenario customizable from the backend. All the researcher needs to do is specify the parameters through a JSON structure (for which we created a template) and upload it to Firebase.

#### 7.3.2.2 Frontend and Interfaces

The next major aspect of NeoCITIES consisted of updating the frontend interface. Figure 7.1 shows what the prior version of NeoCITIES looks like.

Hellar and McNeese (2010) describe the user interface at a high level, and their description is summarized below:

Figure 7.1: Frontend Interface of the Prior Version of NeoCITIES



The resource panel displays the unique set of resources available to each player that can be applied to events. Each resource has a limited number of units that are represented by an icon associated with the resource and a specific badge number. By clicking the Dispatch Units button the interface expands and allows the user to drag and drop resources to selected events on the event tracker.

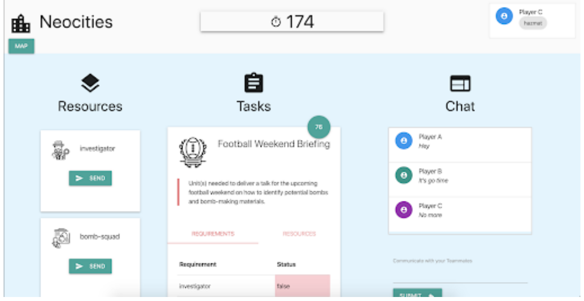
There is also a unit monitor that displays the unit's status (which includes current location and field reports from the event) and enables the player at any time to recall the resource to the home station so that it can be re-allocated. Resources are automatically recalled when an event fails or is completed, but a player can always choose to reassign a resource to a more extreme event.

The redesign of the interface sought to preserve many of the key features (the mechanics of resource allocation, recall, chat, briefs) while also altering it in several ways, mostly to effectively support AI agents.

To start with, Figure 7.2 shows the default view the human participants have access to.

On the right are the resources the player has access to. Players can assign resources through the SEND button, which when clicked turns into a dropdown menu

Figure 7.2: Frontend Interface of the Updated NeoCITIES



Unit(s) needed to deliver a talk for the upcoming football weekend on how to identify potential bombs and bomb-making materials.

| Requirement | Status |
|---|---|
| investigator | false |

Player A
Hey

Player B
It's go time

Player C
No more

Communicate with your Teammates

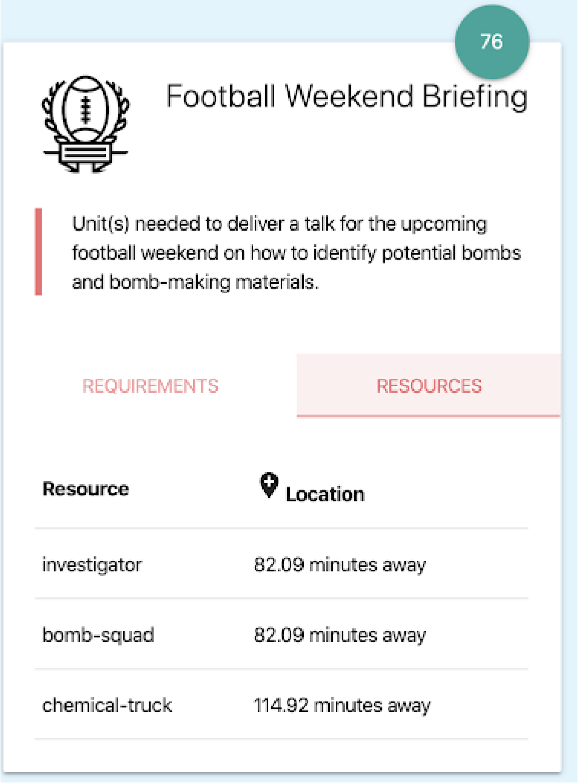
where the user selects which event to assign the resource to. Once the resource is assigned, the button turns into a RECALL button which if clicked recalls the resource to the player's location.

On the left side is the chat component, which identifies each player and their messages through distinct colors. Unlike the prior version of NeoCITIES, the chat automatically scrolls to the bottom so that only the latest 3 messages are immediately visible. If a player needs to go back to the prior message, they just need to scroll upward in the chat. This design choice was implemented because it makes the chat more consistent with how modern applications implement chats, thereby avoiding having the player switch behaviors they've grown accustomed to when using a chat interface.

In the middle are the tasks and events that are currently active. Each event is represented by a card that includes an event description alongside its current status, broken down per requirement (false/red means that particular resource has not arrived to the event yet). The event card also provides two additional types of information, as shown in Figure 7.3.

In the top right it shows how much time is left before the event fails, and if the RESOURCES tab is clicked upon, the card shows the amount of time it would take for each of the player's resources to reach the event. These calculations are done on

Figure 7.3: Event Card



76

Football Weekend Briefing

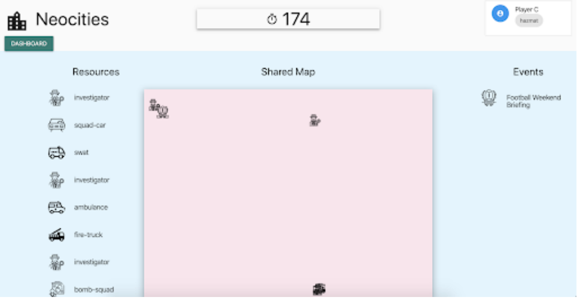
Unit(s) needed to deliver a talk for the upcoming football weekend on how to identify potential bombs and bomb-making materials.

| Resource | Location |
|---|---|
| investigator | 82.09 minutes away |
| bomb-squad | 82.09 minutes away |
| chemical-truck | 114.92 minutes away |

the client side and are updated every time the global state of the simulation changes.

A far larger change however is the implementation of a map locating events and resources, which is shown in Figure 7.4. The events and resources are represented through the same icons that are used on the default view, and are still identified through the legends on the left and right side of the screen.

Figure 7.4: NeoCITIES Map



The last version of NeoCITIES did not have a map, thus the players were forced to coordinate purely looking at the estimates. However, one of the early versions of NeoCITIES did include a map to focus more on shared geo-spatial reasoning among the players. As described above, our version supports both, and specifically includes the map because AI agents can only process states through a matrix representing the coordinates of the elements in the simulation. Thus, were a map not included, the agents would be perceiving the simulation in an entirely different way than the human players, thereby preventing an accurate direct comparison between human-only teams and human-machine teams. Only by having all players have access to geo-spatial information can an effective comparison be drawn.

Overall, participants scroll through and switch from view to view as needed in order to respond to scenarios and its accompanying sequence of events that are being studied.

### 7.3.3 AI Agents and Training

For the first time since its original development, NeoCITIES supports AI agents. These agents are based upon reinforcement learning, a type of machine learning model that learns tasks through behavioral rewards and has recently gained attention for its superhuman performance in Go, Chess, soccer, and Atari games [Hu et al., 1998, Foerster et al., 2018]. Using RL not only enables NeoCITIES to incorporate the most cutting edge AI models, but it also enables the agents to generalize to a wide variety of scenarios since RL agents learn independently from the feedback of the game.

Each agent is an instance of one of the RL models currently existing from both the literature as well as the private sector. To implement the RL agents, we used

Tensorforce, which is an open source RL library focused on providing clear APIs, readability and modularization to deploy RL solutions in both research and real world applications [Schaarschmidt et al., 2017, Schaarschmidt et al., 2018]. There are a host of predefined algorithms present in this library:

- A3C using distributed TensorFlow [Mnih et al., 2016].
- Trust Region Policy Optimization [Schulman et al., 2015].
- Normalized Advantage functions (NAFs) [Gu et al., 2016].
- DQN [Mnih et al., 2013].
- Double-DQN [Van Hasselt et al., 2016].
- Vanilla Policy Gradients (VPG/ REINFORCE) [Williams, 1992].
- Deep Q-learning from Demonstration (DQFD) [Hester et al., 2018].
- Proximal Policy Optimization (PPO) [Schulman et al., 2017]

In order to generate these RL agents, we developed our own training system. RL only works if the agent is able to extensively iterate through the game, thus using NeoCITIES clients and servers would not work because it would take too long. Instead, our training system replicates NeoCITIES mathematically so that each agent can learn through self-play as part of a multi-agent system that encompasses all 3 roles. Specifically, the system represents NeoCITIES as a sequence of matrices of the ever-evolving state of the game. This aspect is critical because the human and the machine in the human-machine teams inherently operate differently (the machine perceives reality through linear algebra representations, the human uses biological perception). The environment is configured to match the scenario that appears on

NeoCITIES' server, thereby removing the need for researchers to create a new environment every time they make changes to the scenario: instead, the researcher simply uses the same JSON file they uploaded to Firebase.

The agent receives the state as a 50x50 matrix with various unique identifiers locating resources and events based on their coordinates. After receiving the state, the agent performs an action by selecting the destination coordinates of each resource at its disposal. The action is processed by the environment, the locations of the resources and the status of the events changes, and the state (which includes the map) is updated. From the updated state, a reward is computed based on how much closer the resource is to an appropriate event compared to its prior state, and the agent then receives the reward. An epoch (an entire session of NeoCITIES) lasts for the time specified in the scenario, but the update happens almost immediately as opposed to taking the interval of time it takes in the real world (the default time interval for updates for NeoCITIES is 5 seconds), which accelerates the agent's training. The agent's model is saved at a regular interval (the default is every 5000 epochs), and continues until a specified amount of epochs have elapsed (the default we used in this study is 200000).

Lasty, we built an API that enables the trained agent to interact with the Firebase database. This API connects to Firebase and generates a game state every time there is an update and polls the agent for an action which is in turn restructured and uploaded to the database. The agent thus responds to a stream of updates coming directly from Firebase, thereby enabling a seamless integration between the human players and the AI agents as NeoCITIES updates on both ends at the same time.

### 7.3.4 Measurements

#### 7.3.4.1 Scoring Model

In order to evaluate human task performance, NeoCITIES relies upon an event growth formula inherited from the original design of the CITIES task [Wellens and Ergener, 1988]. The current formula is designed to incentivize participants to respond to events in a correct and timely manner by allocating resources [Hellar and McNeese, 2010]:

$$M2 = a * M + b * M–C(R)$$

Each event begins with an initial magnitude (M), with the median range of initial values being between 2 and 3. Constants (a, b, c) are the seed coefficients that determine the rate of growth of the event's magnitude (M2), which in turn affects the number of resources (R) required to fully respond. As the number of resources (R) is correctly allocated to the event, the event's severity diminishes proportionally until the event is resolved or it expires because the participants failed to respond in a timely manner [Hellar and McNeese, 2010].

#### 7.3.4.2 Team Metrics

At its core, NeoCITIES tracks performance by evaluating the speed and accuracy of the team's response through raw (cumulative) and relative (weighted) scores. Teams are rewarded for rapidly sending the correct type and amount of resources to an emergency event, and are penalized when an event terminates either because of the team's inaction or through their incorrect or slow response [Hellar and McNeese, 2010]. Table 7.1 matches team metrics with particular measurements computed by the NeoCITIES database.

| NeoCITIES Performance Metrics | |
|---|---|
| Metric | Measurement |
| Task Performance | Raw Score, Relative Score, Response Errors |
| Situational Awareness | Synchronicity, Sequencing |
| Team Communication | Chat Log Frequency |

Table 7.1: NeoCITIES Performance Metrics and their Measurements

Raw score refers to the cumulative magnitude of the events that were successfully responded to by the team. The nature of the scoring model enables the escalation of low-magnitude events into high-magnitude events, challenging teams to effectively prioritize resources between events of varying magnitude. On the other hand, the relative score refers to the raw score being understood as a percentage estimate of possible performance. It's computed by subtracting the ratio of raw score and worst score (the raw score implied by the player taking no action) being subtracted from 100 to retain a positive scale. By weighing events of all severities, the relative score ends up emphasizing performance across all events as opposed to the raw score which is biased towards high initial severity event. Team performance can thus be represented by the average of the teammates' relative scores (highest being best) and the cumulative sum of their raw scores (lowest is best). Response errors refers to the frequency of incorrect resources being allocated to a particular event, and thus functions as a measure of accuracy.

Situational awareness is assessed through two variables: sequencing and synchronicity. The former refers to the degree to which the order in which resources reach an event match some pre-set parameter (highest being best), and the latter refers to the degree to which the correct resources are applied to the event within a narrow time-frame. Both these variables are automatically computed at the end of the session as all interactions between resources and tasks are recorded at the server level.

Lastly, communication is measured in a straightforward way by looking at the message frequency in the chat log. NeoCITIES records and timestamps all communication between players, which can then be analyzed in ways specific to the researcher's design.

### 7.3.5 Study Design

#### 7.3.5.1 Overview

After comparing the performance of three RL agents, we selected the top performing one as the AI agent for the human-machine teams. We then recruited participants on Amazon Mechanical Turk to play NeoCITIES in teams of 3 under different conditions.

#### 7.3.5.2 Task

For the task, we implemented a variation of NeoCITIES default scenario. Each of the three players are assigned to a different role: police, fire, and hazmat. The police role controls an investigator, a squad car, and a SWAT team. The fire role controls an investigator, an ambulance, and a firetruck. The hazmat role controls an investigator, a bomb squad, and a chemical truck. Each player manages these resources to respond to events which become active at different times throughout the session and fail if they are not fully responded to within a certain time (measured in seconds from start time). The list of events is shown in Table 7.2.

The task takes 5 minutes (900 seconds total), and it begins once all players have signed on. At the end of the session, relative score is computed and displayed to the players.

| Title | Resources | Difficulty | Start Time (s) | Time Limit (s) |
|---|---|---|---|---|
| Football Weekend Briefing | investigator | 2 | 11 | 411 |
| Tanker Collision | squad-car<br>fire-truck<br>chemical-truck | 3 | 41 | 441 |
| Escort a Senator | swat | 1 | 126 | 526 |
| Smoking Kills | fire-truck | 2 | 156 | 556 |
| Field Chemical Removal | chemical-truck | 3 | 226 | 626 |
| Luncheon Nausea | Ambulance<br>investigator | 2 | 256 | 656 |
| Possible Student Rave | investigator<br>squad-car | 2 | 329 | 729 |
| Old Main Frame Shoppe Fire | investigator<br>fire-truck | 2 | 359 | 759 |

Table 7.2: NeoCITIES Default Scenario

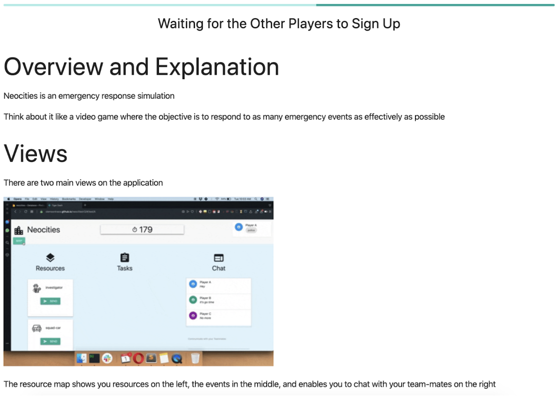
#### 7.3.5.3 Experimental Setup

Each player logs in into NeoCITIES with a different url which identifies the team they are a part of, the session they are about to start (they play a total of 5 sessions), and which role they are assigned (identified through A, B, and C).

Once they login and sign the consent form, they are taken to the training page. The training page includes both a text explanation of how NeoCITIES works, as short videos. At the very top of the page is a progress bar that lets the player know whether all players have logged in. Figure 7.5 shows the training page.

Once all players are logged in, the interface takes the players to the main view and the sessions begins. After the 5 minutes elapse, the relative score is computed and displayed to each player, and a link at the bottom takes them to the next session. Once all 5 sessions are completed, the participants are taken to Qualtrics to fill out

Figure 7.5: Training Page



Waiting for the Other Players to Sign Up

Overview and Explanation

Neocities is an emergency response simulation

Think about it like a video game where the objective is to respond to as many emergency events as effectively as possible

Views

There are two main views on the application

The resource map shows you resources on the left, the events in the middle, and enables you to chat with your team-mates on the right

a survey.

The survey questions extend a basic teamwork principles to include questions drawn from the team cognition literature that identify emergent interaction, shared knowledge, and collaborative environment. These variables are integral to how team cognition is studied in the literature, and emerge from a composite of several survey questions whose cumulative responses have been shown to be correlated with team cognition [Lee and Johnson, 2008].

#### 7.3.5.4 Participants

We enlisted over 60 participants from Amazon Mechanical Turk (MTurk) to play NeoCITIES under different conditions. Traditionally, research with NeoCITIES has been limited to lab experiments because it was developed prior to the advent of cloud computing [Hellar and McNeese, 2010]. Despite initial hesitancy, the use of platforms such as MTurk for behavioral research has grown in popularity among researchers in both psychology and economics [Horton et al., 2011]. The literature has identified the validity-related advantages and drawbacks of using Mturk when compared to traditional lab studies, which serve as a useful starting point to justify our experimental design choice while also discussing potential limitations

[Horton et al., 2011, Paolacci et al., 2010].

To start with, Mturk offers a more diverse pool of subjects to draw samples from. On the other hand, despite its ubiquity, the use of convenience samples in lab studies has recently been heavily criticized as one of the major factors in the replication crisis in experimental economics and psychology [Horton et al., 2011]. Alongside convenience sampling, the heavy reliance on traditional university subject pools by lab studies has also been heavily criticized. MTurk enables researchers to sidestep the problem because the platform's user base has been found to be arguably closer to the U.S. population as a whole than university subjects [Paolacci et al., 2010].

However, prior research has expressed concerns regarding the over-representation of female and Asian subjects in recent experiments relying on MTurk [Eriksson and Simpson, 2010]. The criticism speaks to the distinctions between US Internet users, where an overabundance of female subjects in online study recruitment has been identified, and the US population as a whole [Ipeirotis, 2009, Ross et al., 2010, Gosling et al., 2004]. These concerns, although valid, are reduced in the case of MTurk because the most recent data shows that despite a significant number of workers from India (34%), the plurality of workers is from the US (47%) [Paolacci et al., 2010]. Furthermore, a comprehensive demographic analysis has shown that MTurk workers are at least as representative of the U.S. population as traditional subject pools or more representative than college undergraduate samples and internet samples in general with respect to gender, race, age and education [Paolacci et al., 2010].

Overall, even when directly compared to the kind of lab study NeoCITIES has been traditionally used for, very little evidence exists to suggest that data collected through platforms such as MTurk is necessarily of poorer quality when compared to traditional subject pools [Krantz and Dalal, 2000, Gosling et al., 2004]. Indeed, Horton et al (2011) have shown that MTurk workers engage in the same kind of decision-making that has historically characterized subjects in behavioral research. Additionally, it is in fact that case that Mturk studies has been shown to be far less susceptible to experimenter bias, subject crosstalk, and reactance than lab studies because of the physical separation between the researcher and the research subjects [Paolacci et al., 2010]. These findings imply that MTurk is a very reliable and arguably superior way to conduct research, which alleviates the concern that subsequent studies using our cloud-based version of NeoCITIES won't connect to the literature relying on the prior version.

Weighing these concerns, Horton et al (2011) put forth the following guidelines to ensure that an MTurk study retains validity:

1. Experiments covering the behavior of selected groups (Ex. young mothers) require the use of a contextualized sample for the MTurk subjects to adhere to

2. Experiments involving behaviors in specialized contexts (Ex. after a disaster) also require the MTurk sample to match to the context

3. Estimates of changes (Ex. Do angry individuals take more risks?) are reliable, while estimates of levels (Ex. How many people support candidate X?) much less so

Lastly, the consensus in the MTurk literature heavily emphasise the strong necessity to disclose the demographics of the study's Mturk participants in order to enable differences in the results of similar studies to be compared as to identify the kind of patterns exhibited by cross-cultural studies [Paolacci et al., 2010, Horton et al., 2011]. To that end, we provide a breakdown of our Mturk's samples demographics in the Table 7.3.

|  | Male | Female |
|---|---|---|
| Gender | 67.23% | 32.77% |

|  | African American | Asian | Caucasian | Indian | Other |
|---|---|---|---|---|---|
| Race & Ethnicity | 5.89% | 40.76% | 30.52% | 21.57 | 1.26% |

|  | 18-25 | 26-35 | 36-45 | 46-55 | 56-65 |
|---|---|---|---|---|---|
| Age | 19.45% | 64.31% | 14.81% | 5.03% | 6.40% |

Table 7.3: Demographics

#### 7.3.5.5 Experimental Conditions

NeoCITIES supports 3 players at any given time, which implies 3 combinations of human and AI players: a team of all humans, a team of 2 humans and one agent, and a team of 2 agents and 1 human. We studied each combination as its separate condition. To that end, the participants were grouped in batches of 30, 20, and 10 then assigned to one of the three experimental conditions respectively (human-only team, human-machine team, human-machine-machine team).

In order to select an RL agent to participate as the teammate for the human-machine teams, we trained 3 different RL agents (DQN, PPO, and VPG) and selected PPO to be the teammate in all other conditions (including our benchmark ai-only condition) since it performed the best.

NeoCITIES automatically tracks team performance across the 3 metrics described earlier through the 6 measurements. These metrics were recorded for every round of game play for all conditions, and were combined with the measures from the teamwork principles survey to identify the descriptive ways in which teamwork manifested itself differently across conditions.

## 7.4 Results

We conducted two separate analyses of the results: one of performance metrics, and one of the survey results.

### 7.4.1 Performance Analysis

Our experimental setup is simultaneously between subjects (each team is assigned only one of the team type conditions) and within-subject design (each team, regardless of team type, plays the game 5 times). Pursuant of this dynamic, In order to address how team performance and situation awareness differ across four conditions, we conducted a 4 x 5 mixed Analysis of Variance (ANOVA) to determine whether the different team configurations (four conditions: ai-only, human-ai-ai, human-human-ai, human-only) differed with respect to their performance improvement over time (iteration).

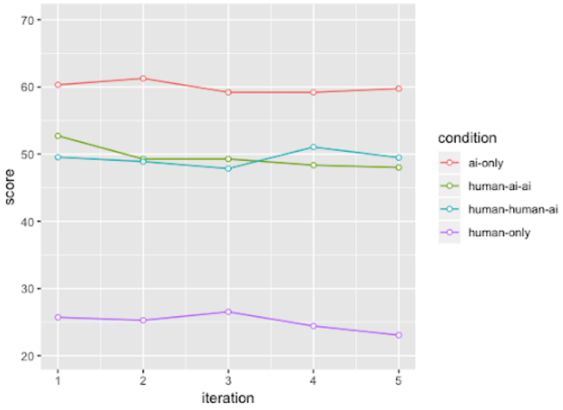
#### 7.4.1.1 Team Score

We began looking at the differences in the scores. As Figure 7.6 shows, there is a wide gap in average performance between the ai-only and human-only teams, while all of the human-machine teams perform in a similar manner and are closer to the ai-only performance.

Breaking down the results in a more descriptive way (Table 7.4), the means of the human-machine teams are extremely close and underperform ai-only teams by 17% (10 points), and the ai-only teams more than double in performance compared to the human-only teams.

The ANOVA analysis displayed in Table 7.5 below shows a few high level patterns:

Figure 7.6: Means Plot of Team Scores



| | Mean | Standard Deviation |
|---|---|---|
| human-only | 25.00 | 3.05 |
| human-human-ai | 49.37 | 5.64 |
| human-ai-ai | 49.53 | 5.65 |
| ai-only | 59.95 | 5.59 |

Table 7.4: Means Table of Team Scores

1. Differences across iterations were not statistically significant (p = 7.99e-02)

2. There is not a statistically significant interaction between condition and iteration (p = 3.52e-01)

3. The differences in performance across conditions were extremely significant (p = 2.25e-28).

Overall, the condition main effect F (3, 36) = 437.83 was the only effect that was significant.

| Effect | SSn | SSd | DFn | DFd | F | p |
|---|---|---|---|---|---|---|
| Condition | 32962.99995 | 903.4338 | 3 | 36 | 437.836158 | 2.257878e-28 |
| Iteration | 77.96752 | 864.6081 | 1 | 36 | 3.246362 | 7.996182e-02 |
| Condition by Iteration | 80.97534 | 864.6081 | 3 | 36 | 1.123867 | 3.522837e-01 |

Table 7.5: Summary of Team Score ANOVA

However, because we did not have a directional hypothesis, we did not interpret the main effects. Rather we performed additional follow-up analyses to better understand the differences between the team types. To that end, we performed a Bonferroni pairwise test, which showed significant differences between all team types.

| | ai-only | human-ai-ai | human-human-ai |
|---|---|---|---|
| human-ai-ai | 5.9e-11 | - | - |
| human-human-ai | 3.6e-11 | 1 | - |
| human-only | <2e-16 | <2e-16 | <2e-16 |

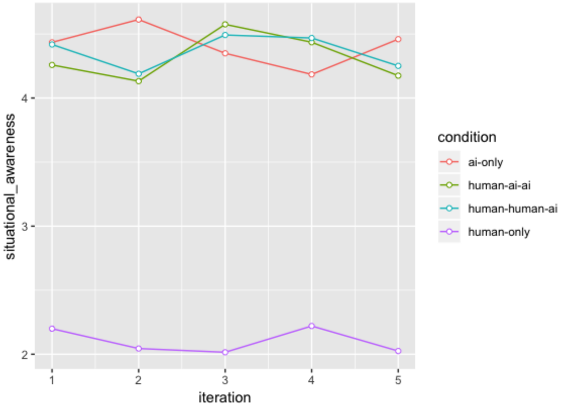
Table 7.6: Bonferroni Pairwise Test of Team Conditions for Team Score

As the Bonferroni comparison in Table 7.6 shows, almost all team configurations differed from each other in a statistically significant way, except for the two human-machine teams, whose differences were minimal to start with.

### 7.4.1.2 Situational Awareness

Next we ran similar tests for situational awareness. Just like with team score, we first plotted the group means of each team condition across all five iterations. The results are shown in Figure 7.7.

Figure 7.7: Means Plot of Situational Awareness



Just as the case of score, the human-only teams substantially underperformed, which makes sense given how situational awareness is highly related to team score (higher situational awareness necessarily means that many resources reached events to complete a task). However, in this case the difference between the human-machine teams and the ai-only teams was much smaller, with the human-machine teams on average outperforming the ai-only teams until the last iteration.

This pattern is more clear once the results are analyzed through descriptive statistics (Table 7.7)

| | Mean | Standard Deviation |
|---|---|---|
| human-only | 2.10 | 0.42 |
| human-human-ai | 4.36 | 0.47 |
| human-ai-ai | 4.31 | 0.54 |
| ai-only | 4.40 | 0.37 |

Table 7.7: Means Table of Situational Awareness

The human-machine teams differ slightly more substantially with each other (human-human-ai teams scored on average 0.32% lower for score vs 1.1% higher for situational awareness), but still exhibit the same situational awareness, which in this case is closer to that of ai-only teams (1% difference in situational awareness vs 1.7% difference in score). Consistent with the gap in team scores, human-only teams displayed under half of the situational awareness of human-machine teams.

The ANOVA analysis (Table 7.8) not only shows that there was not a statistically significant interaction between condition and iteration (p = 8.80e-0), but also that the fluctuations in situational awareness across iterations were not significant (p = 6.11e-01). Much like with score, the only effect that was significant was the condition main effect F (3, 36) = 216.18, p = 4.51e-23.

To better understand the differences between team conditions, we once again ran a Bonferroni pairwise comparison for situational awareness. Table 7.9 below shows the p-values for each pairwise comparison.

| Effect | SSn | SSd | DFn | DFd | F | p |
|---|---|---|---|---|---|---|
| Condition | 191.92161099 | 10.65315 | 3 | 36 | 216.1856620 | 4.507642e-23 |
| Iteration | 0.05534359 | 7.560027 | 1 | 36 | 0.2635400 | 6.107349e-01 |
| Condition by Iteration | 0.13991891 | 7.560027 | 3 | 36 | 0.2220927 | 0.00762372 |

Table 7.8: Summary of Situational Awareness ANOVA

| | ai-only | human-ai-ai | human-human-ai |
|---|---|---|---|
| human-ai-ai | 1 | - | - |
| human-human-ai | 1 | 1 | - |
| human-only | <2e-16 | <2e-16 | <2e-16 |

Table 7.9: Bonferroni Pairwise Test of Team Conditions for Situational Awareness

Unlike with team score, situational awareness is only significantly different when limiting the comparison between the human-only teams and all other team types (p = <2e-16). This finding makes sense given the large gap in situational awareness between human-only teams and the other teams and the far lesser gap between human-machine teams and ai-only teams.

### 7.4.2 Team Cognition Survey Analysis

Next, we shifted our focus exclusively toward the teams that included human participants (thereby excluding the ai-only teams) to analyze the relationship between the performance metrics and team cognition to see if team cognition is a comparable predictor of teamwork for human-machine teams. We accomplished this through a linear regression model that relates team cognition variables to team score as well as situational awareness.

The first team cognition variable we analyzed is team shared knowledge, which is a composite of perception-related sub variables that range from perceived shared knowledge about the task to perceived mutual understanding of individual preferences and communication tendencies. All of these sub variables are elicited through their own specific question in the teamwork principles survey, and have their own designated sub-variable identification. The block of questions for this variable includes the questions between Question #51 ("My teammate has a general knowledge of specific team tasks") to Question #66 ("My teammate strives to express his or her opinion").

Next we analyzed team environment which, much like the prior variable, is a composite of several sub-variables related to trust, perceived rewards tied to behavior, safety, and perceived constraints. All these items were once measured through the survey and aggregated to generate the variable. The block of questions for this variable ranges from Question #67 ("There is an atmosphere of trust among my teammates") to Question #74 ("My team knows the environmental constraints when we perform various team tasks").

The last team cognition variable we analyzed was emergent interaction, which is a composite of mutual role understanding, perceived shared information, perceived interaction level, perceived exchange effectiveness, flexibility, collaborative decision-making, informal communication, and listening. Once again, individual questions designated for each variable in the survey in order to measure these constructs. The block of questions for this variable ranges from Question #42 ("My team understands its roles and responsibilities.") to Question #50 ("My teammates consistently demonstrate effective listening skills").

Lastly, we combined all three variables into one regression model for both team score (score ∼ team_knowledge + team_environment + team_interaction) and situational awareness (situational_awareness ∼ team_knowledge + team_environment + team_interaction). These three variables once combined serve as proxies for team cognition under our methodology. The results are displayed in the Table 7.10.

Our "score" model had a residual standard error of 3.089 on 26 degrees of freedom. Looking the R-squared value, the team cognition variables collectively explain

| Variable | Coefficient | Std Error | P-value |
|---|---|---|---|
| SCORE | | | |
| Intercept | 83.67 | 2.75 | <2e-16 |
| Team Knowledge | -0.42 | 2.75 | 0.07 |
| Team Environment | -1.09 | 0.54 | 0.05 |
| Team Interaction | -1.08 | 0.43 | 0.02 |
| SITUATION AWARENESS | | | |
| Intercept | 7.33 | 0.26 | <2e-16 |
| Team Knowledge | -0.05 | 0.02 | 0.02 |
| Team Environment | -0.05 | 0.05 | 0.29 |
| Team Interaction | -0.11 | 0.04 | 0.01 |

Table 7.10: Team Cognition Linear Regression for Team Score

93.99% of the variance of team scores. The F-statistic is quite large (135.4 on 3 and 26 degrees of freedom) and significant (p = 5.48e-16). Pursuant of these results, it becomes possible to regularize the coefficients and produce beta coefficients that can be more effectively interpreted.

Considering the statistical significance of each team cognition variable, only team interaction has a meaningful influence on team score. Specifically, team score is expected to decrease by 0.46 for every standard deviation increase in team interaction. On the other hand, team score is expected to decrease by 0.26 for every standard deviation increase in team environment however, these findings only neared significance (p = 0.0539) and are not significant enough to be reliable. These results need to be understood in context: the model's intercept is 83.67, thus across conditions it seeks to predict performance losses between human-only and human-machine teams. Therefore, the results do suggest that even though team cognition is more prevalent as more humans become part of the team, performance actually decreases with team cognition because it emerges more intensively among the humans in the underperforming human-only teams than between humans and machines.

On the other hand, our "situational awareness" model had a residual standard error of 0.29 on 26 degrees of freedom. The Multiple R-squared shows that team cognition explained most of the variance in situational awareness (93.91%). As expected, the F-statistic is closely related to that of the “score” model (133.7 on 3 and 26 degrees of freedom, p = 6.42e-16) since, as previously explained. both performance variables are highly related due to the nature of the game. Give the strength of our model, we were able to regularize our coefficients into beta coefficients in this instance as well.

Situational awareness is expected to decrease by 0.36 for every standard deviation increase in team knowledge (p = 0.02). On the other hand, team environment was not statistically significant. Lastly, situational awareness is expected to decrease by 0.49 for every standard deviation increase in team interaction (p = 0.01). Once again, the model’s intercept is 7.33, hence it’s oriented at understanding performance drop offs between human-machine teams and human-only teams, and it does so very effectively.

## 7.5 Discussion

The study enables us to draw inferences from several different perspectives. The different experimental conditions are useful in determining both the way interacting with AI changes human players’ ability to coordinate and adapt to the environment as well as the extent to which team behavior in human-machine teams differs from pure machine teams.

To start with, the survey data analysis enables us to detect the extent to which principles of human-teamwork are as predictive of performance and situational awareness in human-machine teams. This analysis speaks to RQ1, which asks about the applicability of the principles of human-teamwork (specifically, through team cognition) to human-machine teams. Our results run counter to expectations based on extending what the literature as identified as key aspects human teamwork. Indeed, human-machine teams outperformed human-only teams despite lower levels of team cognition.

Specifically, our data indicates that higher levels of team interaction results in lower team scores. One interpretation of this result is that the introduction of an AI into the NeoCITIES team often induces the human players to communicate differently to compensate for the agent's inability to coordinate directly. However, performance is still higher even in teams with only one human, which, in the context of ai-only teams outperforming all other teams, suggests that higher team scores can be achieved by humans following the machine's lead. Still, with respect to team score the other team cognition variables were not significant, which limits the inferences that can be drawn about the relationship between team cognition and team score alone among teams that included humans.

In the context of situational awareness however, the results were more significant. With the exception of team environment, team knowledge and team interaction had a negative influence on situational awareness. Specifically, team interaction twice the impact as team knowledge on situational awareness. One way to consider the result is that the presence of a machine-team-mate who cannot communicate forces the human players to redirect their efforts and more effectively coordinate themselves through NeoCITIES' map, thereby reacting more rapidly as they try to synchronize with the autonomous agent.

Beyond the survey, the performance data speaks to RQ2, which asks in which ways human-machine teams outperform machine-machine teams. Surprisingly, the human-machine teams did not outperform machine-machine teams. As outlined in Chapter 1 and Chapter 2, albeit a limited number, there are strong examples of humans and AIs coming together to outperform AI alone. In our case, our results more closely match those of DeepMind and OpenAI, where over time a RL agent achieves super-human performance through strategies, tactics, and responses that run counter to human intuition.

Given how this is the first time NeoCITIES has been used to study human-machine teamwork, many of the prior results outlined in the literature are thus not as applicable because they only speak to human teams as opposed to human-machine teams. However, one way to frame this finding is by focusing on the unique dynamics of RL. Whereas prior attempts at studying human-machine have suffered from technical limitations (automated as opposed to autonomous agents, wizard-of-oz simulations of agents by humans), it's possible that RL enables the development of agents that eventually get so strong at the task that they can operate almost independently, or are at least more effective as part of a multi-agent system with only AI as opposed to one where humans are also involved.

Lastly, the performance results can be used to address RQ3, which inquired as to whether there'd be performance between different teams of humans and machines. As discussed previously, the ai-only teams substantially outperformed all other team types across all metrics, whereas human-only teams underperformed all other team types. The results also show that the human-machine teams performed closer to ai-only teams (17% lower than ai-only vs. doubling human-only).

Interestingly enough, both types of human-machine teams performed very similarly. While human-human-ai teams scored on average 0.32% lower for team score, their situational awareness was on average 1.1% higher than human-ai-ai teams. These patterns are important because they show that it is not necessarily the case that human-ai-ai teams succeed because of the higher number of agents. Furthermore, the gap between human-machine teams and ai-only teams is far narrower for situational awareness (1% difference) than for scores (17%). This suggests that although the coordination behavior between these teams is very different (for example, human-machine-machine teams cannot rely on communication), both machines and humans are able to adapt to the circumstance to retain high levels of performance and situational awareness. Although they succeed for different reasons, they succeed in much the same way at a complex task such as emergency response management.

## 7.6 Conclusion

In essence, our results suggest that the best way to understand the complex dynamics of human-machine teams is not by attempting to replicate human teamwork dynamics through the design of human-like agents. Rather, a much more fruitful methodology involves leveraging advancements in AI to construct tasks and simulated environments that enable agents and humans to work as peers. Our redesign of NeoCITIES should serve as a useful example on how to accomplish such goals, as well as the value it brings to the research community by making the study of human-machine teamwork more precise and generalizable.

# Chapter 8

# Study 4: Human-Machine Collective Intelligence in Prediction Markets

## 8.1 Introduction

Prediction markets are one of the major settings where collective intelligence takes place. As participants trade shares, forecasting information is aggregated and weighed based on the confidence expressed by the size of each participant's investment. The resulting price thus reflects the market's weighted expectation of the event at a given point in time. By setting up different kinds of prediction markets, it becomes possible to incorporate varying degrees of artificial intelligence to better identify its impact on the overall accuracy of the forecasts produced by each market.

## 8.2 Overview

Prediction markets are one of the major settings where collective intelligence takes place. As participants trade shares in a prediction market, forecasting information is aggregated and weighed based on the confidence expressed by the size of each participant's investment. The resulting price thus reflects the market's weighted expectation of the event at a given point in time. Thus, traditionally prediction markets have been conceived in mechanistic terms. This study however reframes prediction markets as complex systems of humans and machines interacting to better adapt to uncertainty, akin to how teams work together to manage situations.

This study sets up different kinds of prediction markets in order to incorporate varying degrees of AI to better identify its impact on the overall accuracy of the forecasts produced by each market. Specifically, we developed two types of prediction markets: one where bots participate alongside humans (hybrid) and one where only humans participate (human-only). For both prediction markets, the trading data generated by the activity of the participants is used to train several predictive models that are in turn evaluated in their ability to calibrate the prediction market's forecast and make it more accurate.

This methodology thus implements a new prediction market design where AI can play the role of participant (as a bot trading against the humans) as well as that of an aggregator (as a model that uses the price data to refine the forecast). The study then analyzes both prediction markets from a quantitative perspective by looking at the role AI can play (as a participant vs as an aggregator) in enhancing the collective intelligence of the system as measured by its predictive accuracy. The results from the conditions where AI is merely a participant speak to RQ1, while the results where AI functions as the aggregator which processes trading data as its input to then output a calibrated forecast speak to RQ2 and RQ3.

## 8.3 Methods

### 8.3.1 Simulation

Prior research on prediction markets has been limited to the analysis of the performance of prediction markets predicting real-life events as they are about to occur. This approach is limited in several ways: there is no way to measure and track the participants' local knowledge, there is no ground truth that enables the researchers to precisely identify predictive bias, and comparisons between participant conditions are challenging given how the specific events being forecasted only occur once. To sidestep these concerns, we developed a new type of prediction market that enables researchers to effectively simulate events in abstract terms. Specifically, our prediction market operates as an aggregator of local information represented in purely mathematical terms. This methodology simplifies the structure of the prediction market by focusing on the key features of decentralized information gathering and distributed knowledge, and is an extension of Watkins' (2007) methodology [Watkins, 2007]. An example scenario is visualized in Table 8.1.

| | 0 | 0 | 0 | 1 | 1 | 0 | **Belief** |
|---|---|---|---|---|---|---|---|
| Alice | 0 | 0 | 0 | | | | **0** |
| Bob | | 0 | 0 | 1 | | | **0** |
| Charles | | | 0 | 1 | 1 | | **1** |
| Dimitri | | | | 1 | 1 | 0 | **1** |

Table 8.1: Local Knowledge Relative to Ground Truth

Ground truth is represented by the binary vector at the top [0, 0, 0, 1, 1, 0], and each participant has restricted access to a subvector of length 3 [0, 0, 0] and in this case it's assumed they express their beliefs through a simple majority rule. A simple aggregation through this rule produces a 50% ( [1 + 1 + 0 + 0] / 4) probability forecast, which is incorrect because the underlying probability represented by the vector is 33% ( [0 + 0 + 0 + 1 + 1 + 0] / 6 ). Extending Watkins' (2007) analogy, the large vector represents 6 possible qualities that may (1) or may not (0) be present in a candidate's profile, and the participant's vector represents their limited access (only 3 values) to knowledge about the candidate's profile. Even though they are limited in their observation of the candidate's qualities, the participants are tasked to express their beliefs about the candidate's chances at winning the election, and they do by considering whether the majority of the qualities they are able to observe are present (1) or whether they are absent (0), and thus voting accordingly.

However, unlike with traditional prediction markets where the value of the asset being traded reaches $1 if the outcome occurs, or drops to $0 if the outcome fails to materialize, our design ends trading by having the asset reach the actual price as reflected by the ground truth probability (33c in the prior example because the candidate only has 2 out of 6 qualities). This setup still enables the participants to profit by buying and selling shares on the candidate's chances at victory whose price exceeds or is below what they believe to be the correct number.

## 8.3.2 Study Design

### 8.3.2.1 Overview

As mentioned in the previous section, two prediction markets were developed: one where only humans participated, and a hybrid one where humans participated alongside bots. This was done to be able to contrast and identify the ways in which both human and AI behavior change when the two interact. Figure 8.1 provides a

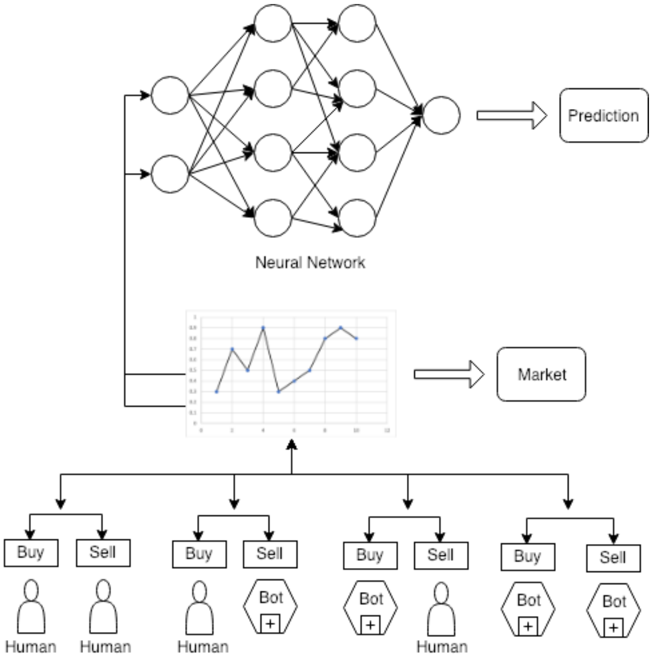


Figure 8.1: The collective intelligence emerging from humans and AIs in the prediction market is fed as input into a neural network

graphical representation of the hybrid prediction market.

Each setup includes a prediction aggregator that takes in the price fluctuations and trading data as its input to output a calibrated forecast that is then evaluated against the ground truth probability. Three types of such aggregators were used (described in the Experimental Conditions section), and they each represent a distinct machine learning paradigm: algorithmic inference, classical methods, and deep learning. Because the aggregators don't influence the prediction market, they can be run independently and simultaneously without compromising the internal validity of each experiment.

#### 8.3.2.2 Task Description

Each asset is represented by a vector of 60 binary features, and a subvector of length 4 is randomly assigned to each participant. The asset's vector represents the ground truth, and the subvector represents the trader's local knowledge. This setup precludes perfect coordination because of the inherent duplicate information (each trader overlaps in knowledge with two other traders), which tasks the participants as well as the aggregator with figuring out how to properly calibrate expected probabilities.

Each participant is given 100 digital tokens at the beginning of the study. These tokens function as the internal currency of the prediction market that enables participants to buy and sell shares of the 10 available assets. Every time a participant submits an order, they are asked to specify a price they are either willing to sell to or buy at, and the order is placed in a queue. In the backend, the queue is divided into batches of 10, and at every iteration buy and sell orders are matched to clear all orders and make sure that as many trades as possible are executed. Every time a buy and sell order was matched, a "trade" occurs, which is recorded as a datapoint on our server along with a timestamp. Once all the possible trades in the batch are executed by the server, the current price of the shares of the asset was updated to reflect the average price at which the trades occurred in the last batch (Ex. 3 trades were executed at 90c and 7 were executed at 80c results in the price to be updated to 3/10 * 90 + 7/10 * 80 = 83c). Participants were rewarded at the very end based on what percentage of the overall available tokens are in their possession.

#### 8.3.2.3 Experimental Conditions

Even though we ran only two prediction markets, our aim is to also understand the usefulness of AI as an aggregator. To that end, we implemented several models (5 to be exact) to compare each type of aggregator (algorithmic, classical machine learning, and neural network) in terms of their ability to improve the probabilistic estimates generated by each prediction market. The models we implemented for each aggregator type are outlined in Table 8.2.

| **Aggregator Type** | **Models** |
|---|---|
| Algorithmic | Linear Regression (LR) |
| Classical ML | Support Vector Machine (SVM)<br>Decision Tree Regression (Tree) |
| Neural Networks | Multi-layer Perceptron (MLP)<br>Long-Short-Term Memory Cell (LSTM) |

Table 8.2: Aggregator Types

Thus, two prediction markets (human-only and hybrid) and the five aggregators (LR, SVM, Tree, MLP, LSTM) became the basis for 10 experimental conditions in a 2x5 design. Through Amazon's Mechanical Turk platform, we recruited 30 participants for each prediction market for a total of 60 participants. For both prediction markets, the participants traded a total of 10 assets one at a time.

### 8.3.3 Measurements

After running the prediction markets, the resulting trading data was collected for analysis and subsequently used as the testing set for the three aggregation methods. At the first stage (prior to the use of aggregators), collective intelligence was measured through different variables: overall accuracy (how close was the final price to the ground truth probability), volatility (how large were the price movements), activity (frequency of trades). At the final state (where aggregators are evaluated), the emphasis will be on the error rate by aggregator, and will be analyzed both at an absolute level (how does the aggregator's estimate compare to ground truth) and at a relative level (to what extent is the aggregator's estimate more or less accurate than the final price). After recording the raw trading data for each of the prediction markets, we structured it in three separate data structures to test which features enable different types of aggregators to be most effective. Then, the price of the event at that time is compared to the ground truth to calculate a bias score, which is what the aggregator is tasked with predicting.

First, we restructured the raw trading data into a 1x30 vector representing how much capital each participant had after every transaction. The second data structure we implemented revolved around price: a 1x2 vector holding the price values for each side of the event (yes, no) at the time. The last data structure combines the capital data structure and the price data structure to enable the aggregator to account for the relationships among the players (reflected in their accumulated capital relative to each other) as well as the current equilibrium among all orders (reflected in the price).

## 8.4 Results

### 8.4.1 Overview

After collecting the transaction data, we generated the datasets through the procedure outlined in the methods section. Every transaction became a data point that included measures from the prediction market as independent variables (structured in one of the several ways outlined in the measurements section) and the difference between the price and the ground truth (hereby referred to as "bias") as the dependent variable.

Subsequently, we proceeded to implement the models to compare their effectiveness as aggregator as measured by their ability to predict bias from the data. Each model was trained on a random selection of 70% of the available data, and tested on the remaining 30%.

For each data structure, we begin the analysis with the algorithmic and classical models: linear regression, support vector machines, and the decision-tree regression. Following the classical models, we implemented two neural networks to identify the predictive capacity of deep learning approaches to our problem. The first deep learning system we implemented is a two-layer multi-layer perceptron (MLP). It received the capital data structure in its input layer, and outputted a float value as its prediction for the bias of the prediction market at a particular time. Since the MLP is a very simplistic example of a deep learning system, we explored the future capabilities of deep learning by implementing a more complex model: long-short-term-memory neural network (LSTM).

### 8.4.2 Price

We began by running the models on the price data version of the dataset, which limits the aggregators to just the price movements for all the shares.

For the human-only prediction market, the results are shown in Table 8.3.

Right away, all the aggregators substantially improve the prediction market's forecast (by at least 67%). The classical machine learning approaches fail to outperform the simple linear regression. Overall, the only aggregator that outperformed the linear regression was the MLP. The MLP's error rate in estimating the prediction

| **Price** | | | | |
|---|---|---|---|---|
| *Model* | *Median* | *Mean* | *Mean - 1std* | *Mean + 1std* |
| Linear Regression | -2.128 | -1.397 | -1.977 | -0.816 |
| SVM | -6.506 | -6.193 | -6.771 | -5.615 |
| Tree | 1.0 | -1.037 | -1.601 | 1.525 |
| MLP | 0.047 | 0.008 | -0.339 | 0.356 |
| LSTM | -33.725 | -33.414 | -33.992 | -32.836 |

Table 8.3: Error Rates of Aggregators Trained on Price Data in the Human-only Prediction Market

market's bias was less than 1%, which gives it a substantial edge over a simple linear regression. However, not all deep learning models outperformed, as the LSTM was the worst performer with an error rate of over 30%.

The results were different for the hybrid prediction market, as shown in Table 8.4.

| **Price** | | | | |
|---|---|---|---|---|
| *Model* | *Median* | *Mean* | *Mean - 1std* | *Mean + 1std* |
| Linear Regression | 0.469 | -0.296 | -1.179 | 0.586 |
| SVM | 0.497 | -0.118 | -0.9993 | 0.763 |
| Tree | 0.100 | 0.094 | -0.857 | 1.047 |
| MLP | -1.073 | -1.301 | -1.975 | -0.626 |
| LSTM | 0.374 | -0.395 | -1.268 | 0.476 |

Table 8.4: Error Rates of Aggregators Trained on Price Data in the Hybrid Prediction Market

At the outset, all models except for the MLP were substantially more accurate in the hybrid prediction market than with the human-only one. However, in this instance it was the LSTM that outperformed all others, potentially indicating that each deep learning aggregator may be better suited for a different type of prediction market.

Once again, the results suggest that the aggregators regardless of type are more effective under the hybrid prediction market condition, suggesting that the introduction of bots trading alongside humans is removing noise from the system which is in turn enabling the aggregation by the AI models to be more successful.

### 8.4.3 Capital

Next, we used the capital data structure as the input for both types of prediction markets. Purely looking at the error rate distributions, no meaningful difference appeared between classical and deep learning approaches. The issue changes once breaking down the basic statistics of each model's predictive success. To that end, we proceed to compare each aggregator through different metrics.

For the human-only prediction market, the results are shown in Table 8.5.

| **Capital** | | | | |
|---|---|---|---|---|
| *Model* | *Median* | *Mean* | *Mean - 1std* | *Mean + 1std* |
| Linear Regression | -2.579 | -2.267 | -2.845 | -1.689 |
| SVM | 6.5 | -6.188 | -6.766 | -5.610 |
| Tree | -2.579 | -2.267 | -2.845 | -1.689 |
| MLP | -33.725 | -33.414 | -33.992 | -32.836 |
| LSTM | -33.725 | -33.414 | -33.992 | -32.836 |

Table 8.5: Error Rates of Aggregators Trained on Capital Data in the Human-only Prediction Market

Under this condition, all methods show substantial effectiveness in estimating the prediction market's bias (lowest level of improvement is 67%), although all of them consistently underestimate by how much the prediction market's forecast deviates from the ground truth. However, neither the classical or deep learning aggregators outperform a simple linear regression, which suggests that machine learning approaches do not add value over a simple regression when trying to calibrate the prediction market's forecast. Specifically, both deep learning aggregators underperform by a large margin (>30% in some cases), thereby reducing the significance of the otherwise encouraging result of a 67% improvement from the raw market prediction at that time.

For the hybrid prediction market, the results are shown in Table 8.6.

| **Capital** | | | | |
|---|---|---|---|---|
| *Model* | *Median* | *Mean* | *Mean - 1std* | *Mean + 1std* |
| Linear Regression | 0.300 | -0.316 | -1.197 | 0.564 |
| SVM | 0.500 | -0.116 | -0.997 | 0.764 |
| Tree | 0.300 | -0.316 | -1.197 | 0.564 |
| MLP | -12.725 | -13.342 | -14.223 | -12.461 |
| LSTM | 0.212 | -0.404 | -1.285 | 0.476 |

Table 8.6: Error Rates of Aggregators Trained on Capital Data in the Hybrid Prediction Market

Under this condition, all aggregators improve substantially when compared to the human-only prediction market. In almost all cases, the aggregators' errors in estimating the gap between the market's prediction and ground truth is less than 1%. The classical machine learning approaches are much closer to the baseline linear regression, whereas the LSTM aggregator outperforms not just in its class (it's the best deep learning aggregator by far), but also overall. The confidence intervals for the aggregators indicate that even in this context, the aggregators underestimate the market's bias far more frequently than they overestimate it.

Once again, the results show that all types of aggregators are more effective under the hybrid prediction market condition, suggesting that the introduction of bots trading alongside humans is removing noise from the system which is in turn enabling the aggregation by the AI models to be more successful.

## 8.5 Discussion

The design of our prediction markets was heavily influenced by Watkins (2007), who identified an effective way to represent local knowledge and collective intelligence through linear algebra [Watkins, 2007]. Thus, our design enabled us to know precisely both the ground truth of the event being forecasted as well as the local knowledge of the participants, which enables a multitude of inferences to be made about the nature of prediction markets and the role AI can play in enhancing their collective intelligence. With respect to RQ1, our results show very strongly that the hybrid prediction market succeeds in making better forecasts than the human-only one, thereby exhibiting higher degrees of collective intelligence. To explain this phenomenon, it's useful to consider the ways in which our design for a hybrid prediction market differs from a traditional prediction market. One of the key differences between a traditional prediction market and our hybrid one is that the latter is designed to stimulate trading without huge costs to the system at the expense of collective intelligence. The prior literature has stressed the obstacles inherent to initial trading in a prediction market before information becomes incorporated in the price [Chen et al., 2010, Hanson, 2003]. Our results show that trading can be catalyzed by the use of trading bots to create enough volatility to incentivize trading without it coming at the expense of any of the human participants.

Beyond stimulating trading however, our results show that introducing trading bots has positive consequences for collective intelligence. The assumption of the literature, influenced by the marginal trader hypothesis, is that players in a prediction market are risk-averse, and thus only expect to profit when the market price deviates form their expectations based on their private beliefs [Wolfers, 2009]. Were that truly the case, then it would not be possible to improve trading by introducing randomized trading bots, yet a few prior studies already show that human forecasting in a prediction market can benefit from the introduction of randomized bots whose trading patterns induce forecasting improvements [Malone, 2018]. Our results go one step further by showing the improvements emerging from the introduction of another type of trading bot that trains itself on the prediction market's trading patterns and trades accordingly.

With respect to RQ2 and RQ3, on the other hand, our results clearly show that collective intelligence can be used as the input to AI so that the latter enhances the former. Specifically, our findings reference one of the core conditions of collective intelligence in prediction markets outlined by Surowiecki (2005): aggregation. The prior literature has primarily focused on the use of different auction structures and matching algorithms depending on the type of forecast that is being elicited from the prediction market [Luckner et al., 2011]. The process of *turning private judgments into a collective decision is still poorly understood.* Chen et al's (2005) attempts at using linear, logarithmic, absolute distance, and quadratic scoring in the context of football forecasts showed very little effect in overall accuracy, whereas our use of classical machine learning models and deep learning have shown impressive levels of improvement in terms of accuracy (at least 70%) [Chen et al., 2005]. However, our results don't suggest that these methods would always achieve these levels of performance in just any prediction market, but rather are tied to the fact that only through our design can the right kind of data and corresponding ground truth be generated through artificial events so that the models can learn and eventually generalize to concrete events being forecasted.

Furthermore, our results speak to another preconditions for collective intelligence to prediction markets according to the prior literature is diversity of opinion [Surowiecki, 2005]. At a technical level, this principle is related to Wolfer's (2009) “marginal trader hypothesis”, where the financial incentive structure of the prediction market incentivizes only traders who believe themselves to possess unique information not shared by the rest of the market. Our results however suggest that in our hybrid prediction market, collective intelligence can emerge despite substantial overlap in participant’s knowledge (each participant in our experiment holds 50% overlap of local information with at least two other players) and belief-less trading (our two types of bots have no access to local knowledge and are merely speculating based on trading patterns).

## 8.6 Conclusion

Overall, the results strongly suggest that the design of a prediction market should not be merely thought of in terms of a mechanism run by matching functions or auction structures, but rather as complex multi-agent systems where humans and AIs can interact to induce emergent properties such as collective intelligence that can in turn become the input to machine learning models to further enhance the forecasting ability of the system.

# Chapter 9

# Connecting The Studies

Given the interdisciplinary nature of the dissertation, it's useful to compare and contrast all of the studies to identify the surprising findings and common themes within the results. By connecting the studies and the models together, it becomes possible to better discern the true nature of human-machine teamwork and human-machine intelligence.

## 9.1 Similarities and Differences Between Study 1 and Study 2

At the outset, Study 1 and Study 2 rely on a similar methodology to compare and contrast how human-machine teams behave differently than either human-only or machine-only teams. Specifically, the game theoretical nature of the experiment makes their results more directly comparable.

### 9.1.1 Cooperation

Both studies' results suggest that different reinforcement learning models exhibit different degrees of cooperation when interacting with humans in strategic game theory scenarios. However, the major point of divergence between them is that in Study 1 cooperative outcomes occur more frequently when the human believes they are playing with another human (even though they are not), while in Study 2 cooperation occurs more frequently when the human participants are indifferent as to whether they are playing against a human or an AI. Essentially, Study 1 suggests that bias towards humans strongly affects cooperation in human-machine teams, while Study 2 shows that the human willingness to cooperate with an AI has only a minor effect on cooperation. Therefore, both studies suggest that the far more influential factor in human-machine teams is the nature of the RL (as characterized by the specific model it operates under) as opposed to the human willingness to cooperate with AI.

### 9.1.2 Coordination

Beyond that however, the results of Study 1 also include an aspect entirely not considered by Study 2: coordination. While Study 1 and Study 2 both study cooperation, understood as the ability for the human-machine to reach the mutually beneficial outcome despite their individual self-interest to the contrary, Study 1 also studies what happens when human-machine teams are tasked with coordination, understood as the ability for the human-machine team to agree and match in a decision over multiple mutually beneficial outcomes. Specifically, the data from Study 1 shows that in a more complex game theoretical scenario, the human-machine team is not only more cooperative, but it also converges to the outcome that most benefits the human far more frequently than to the outcome that benefits the AI the most.

Furthermore, Study 1 shows that coordination, much like cooperation, is strongly mediated by the type of RL agent that is part of the human-machine team. In that respect, the studies both suggest that different RL agents behave differently in a human-machine team, both when it comes to cooperation as well as coordination.

### 9.1.3 Applicability of Human-Teamwork Principles

On the other hand, the major aspect addressed by Study 2 but not Study 1 is the extent to which principles of human teaming are effective predictors of cooperation. Specifically, Study 1 showed that there is a substantially significant relationship between perceived shared understanding and cooperation. Thus, Study 2 and its results will serve as a point of comparison for the other studies (especially Study 3) with respect to the relationship between the principles of human teamwork and human-machine teamwork.

### 9.1.4 Implications

Overall, both studies indicate that the type of RL agent has a strong influence on both coordination as well as cooperation. This implies that researchers should not simply assume that the results from one RL model generalize to all models, which is especially relevant in the context of AI-safety: all models should be tested to ensure safety in human-AI interactions. Furthermore, shared awareness is the far stronger predictor of human-machine teamwork when compared to human player's attitudes towards AI and other key principles in human teamwork. This implies that in multi-agent systems and human-machine dyads, the principles of human teamwork are not as applicable or useful as they are with human teams.

## 9.2 Similarities and Differences Between Model A & Model B

The two models not only helped inform the design of Study 3 and Study 4, but they also serve as useful theoretical contributions in their own right. Specifically, they help contextualize the methodologies of Study 3 and Study 4 by situating them within a larger research context so that the results can crystallize the connections between the variety of cognitive models and teamwork models referenced throughout this dissertation.

### 9.2.1 Crowd vs Team

Model A focuses on identifying the features of collective intelligence as well as the behavioral biases in prediction markets highlighted in the literature in order to design a prediction market that addresses such biases while still preserving the properties that enable collective intelligence to emerge. Model B on the other hand extended the implications of the Superforecasting findings in order to update our understanding of both team cognition as well as collective intelligence. In the process, it produced a new model to design teams with. Thus, even though they both speak to collective intelligence, at a basic level Model A is focusing on optimizing a pre-existing structure to enhance collective intelligence in crowds, while Model B brings forth a new model to think about teams by synthesizing several perspectives in cognitive science.

### 9.2.2 Independence vs Feedback

Whereas Model A focuses on ensuring strong levels of independence among all the components of its model in order to preserve collective intelligence, Model B focuses far more on the design of a feedback loop system between teams and AI. This distinction becomes self-evident when considering how the aggregator's estimates in the prediction market at the heart of Model A are not available to the traders, whereas they are in the case of Model B. Model A considers the prediction market as a collective intelligence mechanism primarily concerned with generating the input for an AI, while Model B considers the human team and the AI as a feedback loop aimed at enhancing the intelligence of the overall system by improving all of the components.

### 9.2.3 Roles

Model A and Model B share similarities in the role they assign for AI. Both models integrate AI as an aggregator of individual opinions insights by the participants, in the hope of leveraging machine learning to better understanding how to combine the team or the group's insights and calibrate predictions. Model A however goes one step further and enables AI to participate alongside the humans as a peer by having different bots trade in the prediction market to improve humans' trading behavior.

### 9.2.4 Implications

Model A grounds Study 4 in many different research domains. It identifies the core features of collective intelligence in order to develop a more abstract view of collective intelligence that can incorporate AI. It also outlines the behavioral biases that have plagued prediction markets and have limited their ubiquity in all major forecasting enterprises despite their potential, and it does so in a way that generates precise targets for improvement. Subsequently, model A also provides a deep technical understanding of the mechanics of prediction markets, which is a necessary precondition for Study 4's ability to address the aforementioned behavioral biases and thereby increase prediction market's forecasting accuracy.

Model B on the other hand shows through analysis of the Superforecasting results that team cognition and collective intelligence may be manifestations of the same emergent cognitive phenomenon. This insight lies at the foundation of the entire dissertation, which seeks to develop a nuanced understanding of human-machine intelligence and its emergence at multiple scales.

## 9.3 Similarities and Differences Between Study 1 & 2 and Study 3

Study 1 and Study 2 looked at cooperation as a proxy for human-machine teamwork, whereas Study 3 looked at NeoCITIES's performance metrics (team score and situational awareness) as the teamwork measure. Despite the major difference in experimental setting (Study 1 and Study 2 rely purely on game theory, which implies a static environment, whereas Study 3 uses NeoCITIES, a sophisticated simulation that makes the environment dynamic), they both showed team-level coordination can occur in human-machine teams despite the team's inability to verbally communicate.

### 9.3.1 Performance

The major difference however lies in the fact that in Study 3 the performance difference between human-only and human-machine teams was far larger than it was in Study 1 and Study 2. Furthermore, Study 3 offers a more accurate point of comparison between human-only and human-machine teams because the human-only condition in the experiment included actual humans playing with each other, as opposed to a reverse-Turing condition such as in Study 1 and Study 2, which at best serves as an indirect proxy of the game theoretical decision-making of human-only teams. On the other hand, Study 1 and Study 2 offer more effective comparisons between machine-only and human-machine teams, thereby giving us better insights into how AIs change their behavior when playing against a human as opposed to another AI. In that respect, Study 3 provides better results that speak to the reverse I.E. the extent to which humans engage in more effective forms of teamwork with other humans as opposed to AIs.

Specifically, a major divergence occurs when contrasting the three studies on whether the human-machine teams deviated substantially from baseline behavior. Study 1 and 2 show sharp deviations from the equilibrium of the game among human-machine teams, especially when the human participants believed themselves to be playing with an AI. Study 3 on the other hand showed comparatively more change in performance between human-only and human-machine teams than between human-machine and machine-only teams. Essentially, the difference is smaller at the smaller scale of a dyad than it is at the larger scale of a triad.

### 9.3.2 Team Structure and Team Dynamics

The difference in team structure is also important. Study 1 and Study 2 provide important results for dyads in a setting that include adversarial dynamics. Specifically, in all three game theory scenarios, the humans and AIs have to reconcile their individual self-interest with that of the two-agent team. This dynamic creates the incentive to betray one another, and thus speaks to trust. Study 3 on the other hand produced results for a triad in a setting where the only objective was a shared performance goal. The lack of incentives to betray one another makes the results between Study 1, 2, and 3 stand out in sharp contrast. Thus, the fact that the studies have different results when it comes to differences in performance speaks to how human-machine teamwork manifests itself differently between dyads and triads.

### 9.3.3 Applicability of Human-Teamwork Principles

Furthermore, Study 2 and Study 3 both look at the extent to which principles of human-teamwork were predictive for human-machine teamwork. The results of Study 2 show that in a two-way interaction between humans and AIs human teamwork principles, especially shared awareness, are strong predictors. Study 3 instead shows that many of those same principles, team cognition specifically, are weak predictors of human-machine teamwork with respect to team score.

### 9.3.4 Implications

This contrast suggests that when transitioning into more complex settings and teamwork structures, the usefulness of traditional models from human teamwork in predicting human-machine teamwork success breaks down. Human-machine teams need to be understood on their own terms, integrating both the humans and the machine perspective. In the case of the latter, multi-agent systems are a useful starting point because, as the results suggest, incentive structures strongly influence cooperation and coordination in RL agents.

## 9.4 Similarities and Differences Between Study 3 and Study 4

Study 3 and Study 4 approached human-machine intelligence from two different perspectives. Whereas Study 3 used NeoCITIES to study human-machine intelligence at the micro-level, Study 4 used a prediction market study it at the macro-level. Specifically, the results of Study 3 speak to human-machine intelligence from a more traditional team perspective, whereas Study 4 approaches it from a complex system perspective.

### 9.4.1 Team Structure and Team Dynamics

Furthermore, these results need to be understood in their proper context. Study 3 is looking at human-machine teamwork manifested in a team's ability to effectively adapt and respond to changes in a physical situation. Study 4 on the other hand looks at human-machine teamwork in a complex system via a prediction market, thus human-machine teamwork manifests itself in the overall system's ability to make accurate forecasts under deep uncertainty. To reframe the contrast in a different way, Study 3 shows that human-machine teams respond to a dynamic physical environment in a manner akin to that of human-only and machine-only teams, whereas Study 4 shows that hybrid prediction markets are more accurate in assessing risk and uncertainty in a probabilistic setting.

One of the major implications of these two approaches is that unlike Study 3, Study 4 examined the role AI can play in human-machine intelligence through different roles. Whereas the members of the teams in Study 3 all operated as peers, AI in Study 4 played the role of both a participant as well as aggregator. Specifically, Study 4 shows that not only is the collective intelligence of the prediction market increased when AI participates and trades alongside humans, but that even when being relegated merely to the role of aggregator major improvements still occur. Therefore, the results of Study 4 speak to the importance of the different roles AI can play as well as to the value of effective role design in successful human-machine teamwork.

### 9.4.2 Types of Emergent Cognitions

Another major distinction between the two studies is the nature of the inference drawn from the results. Specifically, Study 3 looks at the correlation between team cognition and performance to determine whether team cognition was a predictor in performance, and thus the survey answers do not affect variables such as score. On the other hand, in Study 4 collective intelligence is directly related to improvements in forecasting accuracy because the former is framed in terms of the latter, and thus improvements in collective intelligence have a causal relationship to improvements in forecasting. Therefore, Study 4 shows that collective intelligence is a much better predictor of human-machine teamwork than team cognition.

### 9.4.3 Performance

However, the major similarity between Study 3 and Study 4 lies in the convergent results. Specifically, Study 3 shows human-machine teams to be more far more successful than human-only teams with respect to team cognition and performance, and the results in Study 4 show major improvements in hybrid prediction markets compared to human-only prediction markets. Not only is the hybrid prediction market more accurate than the human-only prediction market, but all types of aggregators more effectively correct for bias when trained on the hybrid prediction market's data. Furthermore, the human-machine results in Study 4 stand in sharp contrast to those of Study 3 in much the same way as Study 3's results deviate substantially from those of Study 1 and 2.

### 9.4.4 Implications

As mentioned previously, the difference between human-machine performance in Study 4 and in Study 3 is comparable to the difference between the latter and Study 1 and 2. The pattern in the results of human-machine performance is that it rises with the scale of human-machine interaction (multi-agent dyad, human-machine triad, hybrid complex system) and the complexity of the task (overcoming game theoretical dilemmas, coordinating a response to an emergency, forecasting under uncertainty).

## 9.5 Similarities and Differences Between all Studies

All the studies approached human-machine teamwork and human-machine intelligence from different perspectives. Study 1 and Study 2 used game theory to study humans and machines working together as a multi-agent system. Whereas Study 3 used NeoCITIES to study human-machine intelligence in triads (as opposed to dyads like in Study 1 and Study 2), Study 4 used a prediction market study it at the macro-level as a complex system.

### 9.5.1 teamwork without Communication

The first major insight shared by all studies is that human-machine teamwork can occur despite the inability for humans and AIs to directly communicate. This stands in sharp contrast to human teamwork, which heavily relies on communication to enable coordination among the teammates. Indeed, despite the fact that in none of the studies was communication between humans and AIs was possible, human-machine teamwork still emerged. Specifically, the performance of human-machine teams was always substantially better that of human-only teams (in the case of Study 3), comparable (such as in the case of Study 1 and Study 2), or vastly superior (such as in the case of Study 4).

### 9.5.2 Cooperation and Teamwork

Second, given the ability for humans and AIs to cooperate over shared goals, cooperation often translates into teamwork, but not as understood in human terms. Study 1 and Study 2 show that humans and AIs can cooperate despite incentives not to do so, however Study 3 suggests that team cognition dynamics do not necessarily translate into more effective teamwork in a more complex setting. Furthermore, Study 4 shows that when the interaction between humans and AIs is framed in almost purely adversarial terms (trading on the prediction market is a zero-sum game), collective intelligence still emerges, and is in fact higher than the collective intelligence emerging from a prediction markets where only humans participate. However, Study 4 also shows that when humans and AIs are ensembled (since the prediction market's prediction is combined with that of the aggregator), collective intelligence also improves.

### 9.5.3 Differences at Scale

Another way to understand this difference is by framing the results of each study in terms of the size of the human-machine teams. Specifically, the combined results of all of the studies show that human-machine teamwork is sensitive to scale. Study 1 and Study 2 show that dyads can effectively cooperate to overcome negative outcomes from prisoners-dilemma type games. Study 3 however shows that with teams of three players, human-machine teams don't perform substantially differently from each other but they do perform better than human-only teams. Study 4 on the other hand shows that at a much larger scale (I.E. that of a crowd in a prediction market), humans and machines can work together in ways that produce higher degrees of collective intelligence. Therefore, the emergent phenomena of human-machine teams change as the size of the human-machine team increases.

### 9.5.4 Implications

When compared to all the other studies, the results of Study 4 suggest that the model that best relates to human-machine teamwork is that of collective intelligence. Study 1 and Study 2 approach human-machine teamwork from the perspective of a multi-agent system; however, the essential prediction under that framework would be for the human-machine team to converge towards the Nash Equilibrium, which our results show was not the case. Study 3 approaches human-machine teamwork from the perspective of team cognition; however, our results show that not only do human-machine teams not perform differently than other types of teams, but also that team cognition is not a strong predictor of human-machine team performance. Contrasting that to the results of Study 4, which show major improvements in forecasting once AI entered the picture (as an aggregator, participant, or both), it's clear that many of the principles of collective intelligence apply to human-machine teams and can thus be used to better design prediction markets.

# Chapter 10

# Conclusion

In Chapter 1, several objectives were outlined for this dissertation. Specifically, the following goals were articulated:

1. to identify the similarities and differences between human-machine teamwork, human-human teamwork, and machine-machine teamwork;
2. to explore the unique ways in which team cognition emerges in human-machine teams;
3. to identify the ways in which AI can enhance collective intelligence in human-machine teamwork;
4. to develop empirically-backed design guidelines for the integration of AI in prediction markets

This dissertation sought out the goals of studying and connecting multiple aspects of cognitive science and artificial intelligence by addressing many of the poorly understood connections between team cognition and collective intelligence. Through the four studies, the goals outlined in Chapter 1 have been met.

## 10.1 Research Questions

Specifically, the results now enable me to answer all of the research questions in ways that contribute to multiple research communities. Below the results will be outlined in terms of how each research question specifically.

### 10.1.1 RQ1: Which principles of human teamwork are applicable to human-machine teams?

The main studies that addressed RQ1 were Study 2 and Study 3. Study 2 looked at shared understanding, optimism towards artificial intelligence, and preference towards human teammates. Study 3 looked a several of the variables that constitute team cognition, such team interaction, team knowledge, and team environment.

The main takeaway from Study 2 was that shared understanding was a major predictor of the human-machine team's cooperation levels in a game theoretical dilemma. On the other hand, team satisfaction and perceived trust were not significant predictors, thereby showing that not all principles of human teamwork translate effective in the human-machine setting.

Study 3 on the other hand focused almost entirely on team cognition. The hypothesis was that team cognition would serve as a good predictor of human-machine team performance, yet human-machine teams outperformed human-only teams despite lower levels of team cognition. With respect to team scores, our data indicates that higher levels of team interaction results in lower team scores. However, the other team cognition variables were not significant, which limits our ability to draw meaningful inferences about the relationship between team cognition and team score in human-machine teams. With respect to situational awareness on the other hand, team interaction had twice the negative impact as team knowledge on situational awareness.

Therefore, only a few of the team cognition variables can serve as useful performance predictors, but they are all negatively correlated with performance and thus display the inverse relationship they have with human teamwork. Comparatively speaking, performance, whether in the form of team score or situational awareness, is far more influenced and predicted by the agent's RL model (Study 2) or the number of machines in a team (Study 3) than by any of the principles the literature associates with successful human teamwork.

### 10.1.2 RQ2: In what ways do human-machine teams outperform human teams and machine-machine teams?

Going into this dissertation, the expected outcome was that human-machine teams would consistently outperform human-only and machine-machine teams. This was the case in Study 1 and Study 2, but only somewhat in the case of Study 3.

Study 1 showed that cooperation is more successful in human-machine teams because machine-machine teams converge more often on the sub-optimal and mutually harmful Nash equilibrium. The results of Study 2 match this pattern as well, with the human-machine teams engaging in longer periods of peace and cooperation. However, Study 1 also showed not only that human-machine teams coordinated more effectively than machine-machine teams, but also that they were consistently more likely to converge towards the mutually beneficial outcome that benefited the human the most as opposed to the one preferred by the machine.

Study 3 on the other hand showed human-machine teams only be outperforming human-only teams, but not ai-only teams. Although surprising, this result can be understood in light of the major strengths of the contemporary RL models used to develop the NeoCITIES agents. As outlined in Chapter 1 and Chapter 2, RL agents have recently achieved super-human performance across a variety of domains; thus, the examples of humans and AIs coming together to outperform AI alone we previously referenced can be understood as exceptions to a much larger pattern of agents eventually getting so strong that they perform best either independently or as part of a machine-machine team.

### 10.1.3 RQ3: What are the performance tradeoffs between human-human, human-machine, and machine-machine teams?

The results of Study 3 pointed to two extremes: on the one hand, the machine-only teams substantially outperformed all other team types, while on the other the human-only teams substantially underperformed. This was true across all team types (the machine-only teams outperformed the human-machine teams, while the human-only teams underperformed) and across all metrics (the same results held for both team score and situational awareness).

Beyond the machine-only and human-only gap, both types of human-machine teams performed very similarly. Furthermore, the results show that the performance of the human-machine teams was closer to that of machine-only teams than that of human-only teams. Specifically, the team scores of human-machine teams were on average only 17% lower than those of the machine-only teams while also being over double those of human-only teams. However, we cannot infer that such results for the human-machine teams are attributable to the higher number of RL agents. This becomes clear when considering the performance differences between human-human-ai teams and human-ai-ai teams: on average, the former's team scores were 0.32% lower than the latter while situational awareness was 1.1% higher. Albeit small, these differences were statistically significant, and thus give us sufficient evidence to reject the notion that human-machine teams approach machine-only performance purely because of the presence of RL agents on the team.

Lastly, although the coordination behavior in the human-machine teams is very different (communication is not possible in the case of human-machine-machine teams), the teams are able to adapt and obtain high team scores and retain higher levels of situational awareness. Interestingly enough, the gap between human-machine teams and ai-only teams for situational awareness (1% difference) is far narrower than that for team scores (17%).

### 10.1.4 RQ4: Do human-machine multi-agent system exhibit higher degrees of collective intelligence than human-only systems?

This question was addressed in a straightforward manner through Study 4, which reframed prediction markets as human-machine multi-agent systems. RQ4 is answered affirmatively in multiple ways. First, at a basic level the hybrid prediction market outperformed the human-only prediction market, thereby showing a higher level of collective intelligence. Second, when using a machine learning model as the aggregator, both prediction markets became more accurate, showing that even when limited to a more external role, the human-machine ensemble outperformed. Lastly, across all models used to test different aggregation methods, the hybrid prediction market was most successful, indicating that enabling AI to participate as a trading bot alongside the human participants generated the kind of data that enables all types of aggregation to be more accurate.

### 10.1.5 RQ5: Can collective intelligence be used as the input to artificial intelligence?

Our results in Study 4 clearly show that collective intelligence can be used as the input to AI so that the latter enhances the former. Specifically, our findings reference one of the core conditions of collective intelligence in prediction markets outlined by Surowiecki (2005): aggregation. The prior literature has primarily focused on the use of different auction structures and matching algorithms depending on the type of forecast that is being elicited from the prediction market [Luckner et al., 2011]. The process of turning private judgments into a collective decision is still poorly understood. Chen et al's (2005) attempts at using linear, logarithmic, absolute distance, and quadratic scoring in the context of football forecasts showed very little effect in overall accuracy, whereas our use of classical machine learning models and deep learning have shown impressive levels of improvement in terms of accuracy (at least 67%).

However, our results don't suggest that these methods would always achieve these levels of performance in just any prediction market, but rather are tied to the fact that only through our design can the right kind of data and corresponding ground truth be generated through artificial events so that the models can learn and eventually generalize to concrete events being forecasted.

### 10.1.6 RQ6: Which machine learning approaches are best suited for prediction markets?

Study 4 was designed to answer this question at multiple levels given how AI permeates each of the experimental conditions in different ways. Specifically, with both types of prediction markets the data was tested as the input to a machine learning model which was in turn benchmarked against a simple linear regression.

In the case of a human-only prediction market, the 2-layer neural network (MLP) trained on the price data substantially outperformed all other methods: its error rate was 0.008, which is substantially better than the second best which was the decision tree also trained of price data (-1.037 error rate). Furthermore, Study 4 shows that in the case of a human-only prediction market, all aggregation methods perform worse after being trained on the capital distribution as opposed to the price data. In the case of the hybrid prediction market however, it was the classical machine learning approaches (decision tree and SVM) that outperformed all others. Specifically, a decision tree trained on price volatility performed 19% better than the two closest models, which were the SVMs trained on either data structure.

Overall, the top performing model was the MLP (0.008) trained on price data from the human-only prediction market , followed by a decision tree (0.094) trained on the price data from the hybrid prediction market and an SVM (-0.116) trained on the capital distribution data from the hybrid one . One way to interpret the result is that training deep learning systems on human-only prediction markets achieves the best accuracy. However, the gains are somewhat marginal: the much larger and significant pattern is that implementing any machine learning model and training it on either the price data or the capital distribution leads to massive improvements in accuracy. This insight becomes self-evident when considering that the lowest level of forecasting improvement by any model in either prediction market was 67%.

Furthermore, fixating on the marginal gains between models distracts from the much more important observation that with the exception of one instance of the MLP, all models improve when trained on the hybrid prediction market's data. Therefore, most of the gains are achieved by integrating machine learning both as a participant through the trading bots and as an aggregator.

## 10.2 Overall Study Contributions

When combined, the studies and models making up this dissertation lead to several major implications for research at the intersection between cognitive science and computer science, team cognition and collective intelligence, and multi-agent systems and prediction markets.

### 10.2.1 We should not limit our understanding of human-machine teamwork to the principles of human teamwork

Given the technical limitations that have historically affected the development of AI agents, the research community's understanding of the difference between human teamwork and human-machine teamwork has been severely limited. This dissertation is the first attempt at using reinforcement learning to study human-machine teamwork by leveraging a platform like NeoCITIES, which has been validated to be an effective tool for analysis for human teamwork. As mentioned previously, the differences in scores and situational awareness between the two types of human-machine teams in Study 3 (human-machine-machine and human-human-machine) suggest that their superior performance compared to human-only teams cannot be attributed merely to the presence of the agents, but rather emerge from dynamics not yet fully understood. Given the long history of team cognition effectively predicting human teamwork in NeoCITIES, the fact that the team cognition model is not a meaningful predictor of human-machine performance undermines the notion that human-machine teamwork can be fully understood through the lenses of human teamwork.

Furthermore, this dissertation as a whole makes a compelling case that there are far better models to understand human-machine teamwork: multi-agent systems and collective intelligence. The former is useful to understand the pivotal role incentives and game theoretical dynamics play in cooperation between humans and machines. Specifically, Study 2 shows that that most principles of human teamwork, even something that is usually thought of as critical as communication, are poor reference points in understanding how humans and machines can work together to accomplish goals. However, Study 1 and Study 2 also show the limitations of the multi-agent system model, because the humans and the RL agents in the experiments often managed to escape the mutually harmful Nash Equilibrium to successfully cooperate despite the incentive not to do so. In turn, these results show that game theory alone cannot fully explain human-machine teamwork, which would otherwise be the case if human-machine teamwork could be entirely reduced in terms of a multi-agent systems model.

Study 4 on the other hand shows how collective intelligence is a useful model to understand the impact of human-machine interactions on the intelligence of a prediction market. The results also show several different ways in which collective intelligence emerges through different types of relationships between humans and machines: horizontal, in the case of the machine fulfilling a highly differentiated role as a trading bot, and vertical, in the case of the AI as the aggregator combining the data generated by the prediction market and separately calibrating predictions. However, just like with the multi-agent system model, the collective intelligence model also has limitations in its ability to effectively account for human-machine teamwork. Specifically, the methodology used in Study 4 shows how having an AI play the role of aggregator enables the emergence of collective intelligence in the prediction market despite substantial overlap (50%) in participant knowledge, which runs contrary to the prior literature's traditional understanding of how collective intelligence emerges in crowds. Indeed, this dissertation shows how the introduction of AI can enable humans to transcend many of the behavioral biases that plague prediction markets, thereby demonstrating how collective intelligence is not tied the marginal trader hypothesis.

Thus, it may be more productive to move away from researching human-machine teamwork through analogies to human teamwork, and instead engage in a more interdisciplinary approach that leverages technical advances to not just explore the connections between machine-machine teamwork and human-machine teamwork, but also between multi-agent systems and prediction markets.

### 10.2.2 Team cognition and collective intelligence are analogous phenomena happening at different scales

Another major contribution of this dissertation is the thorough analysis of the team cognition and collective intelligence literature to classify similarities and differences. Specifically, this dissertation uses the human-machine teamwork setting to show how both models speak to a similar emergent phenomenon albeit a different scales. While team cognition explains the emergence of a shared mental model as teammates respond to a physical situation, collective intelligence speaks to the emergence of higher-level information processing as a crowd navigates a probabilistic setting. In both cases, a new and different cognitive process arises in the team that cannot be accounted for at the individual level, but that share many of the same properties that result in higher degrees of performance.

Overall, despite their differences, team cognition and collective intelligence have many core features in common. This dissertation builds upon this insight by showing how the Superforecasting research can help us reframe team cognition and collective intelligence as manifestations of the same cognitive phenomenon. Specifically, researching Superforecasters shows how a large crowd is no longer necessary to have collective intelligence as long as teams are trained through particular cognitive strategies when engaging in forecasting. Additionally, this dissertation's analysis of Superforecasting opens up a new research landscape for team cognition by showing how team cognition is not limited to perception, military settings, and responses to physical situations, but can also occur in prediction, forecasting, and probabilistic decision-making.

In essence, this dissertation suggests that a new type of intelligence can be unlocked through teamwork that goes beyond what the team cognition literature has been demonstrating. By identifying the analogous nature of the two phenomena, collective intelligence expands ways of thinking about team cognition beyond the situational response setting, and team cognition deconstructs the prevailing notion that collective intelligence is merely the by-product of the law of large numbers in a sample of random guesses. Specifically, as a team of Superforecasters displays both team cognition and collective intelligence, the two emergent phenomena, thought to be separate up until that point, can be unified as complementary aspects of a higher-level form of information processing that in turn can become the foundation for human-centered technology designed to enhance teamwork.

### 10.2.3 Thinking of prediction markets as human-machine teams improves collective intelligence by enabling the effective use of AI

As mentioned previously, the current literature on prediction markets is almost entirely focused on a mechanistic view. To that end, most literature on the design of prediction markets is focused on improving the efficiency of the matching algorithm, on structuring different types of auctions, or on the most effective way to model real world events. This overly technical perspective comes at the expense of asking deeper questions about the nature of collective intelligence and the dynamics underlying its emergence in prediction markets.

This dissertation stands in sharp contrast to that approach, and shows how bridging the gap between the communities of cognitive science and computer science is critical in order to design more effective prediction markets. Specifically, Study 4 shows through both its methodology and its results how to conduct more robust experiments with prediction markets, which in turn supports new research validating teamwork approaches to the design of new and better prediction markets. Through this framework, a prediction market thus becomes a larger form of a human-machine team where multiple humans and multiple machines work together to generate the data necessary for the aggregator to improve the system's collective intelligence.

In essence, by thinking of collective intelligence as a macro-form of team cognition, we were able to design a new type of prediction market that leverages both human as well as machine intelligence to aggregate local knowledge effectively to produce very accurate estimates of probabilistic outcomes. Our design shows that AI can play a role in enhancing collective intelligence by both participating and trading alongside human participants to nudge them towards more effective forecasting, and also as the overall aggregator of the information generated by the prediction market in order to calibrate its predictions and thus offset many of the behavioral biases identified in the literature.

## 10.3 Intellectual Contributions to Academic Communities

Given the interdisciplinary approach of this dissertation, the academic contributions span several domains. Specifically, the studies and the models combined reference major concepts from cognitive science, game theory, human factors, human-computer interaction, and machine learning. As a result, we organized the implications and contributions of this research into three major areas: artificial intelligence, team cognition, and collective intelligence.

### 10.3.1 Artificial intelligence

As discussed in Ch 1, most artificial intelligence research right now is being conducted with the aim of replacing as opposed to augmenting human effort. This dissertation shows several ways of keeping human effort in the loop by validating the need to study human-machine teamwork as AI becomes more sophisticated.

Specifically, as reinforcement learning evolves, RL agents should be studied in a human-machine team setting as well in order to better understand the complex dynamics of these autonomous agents interacting with humans in the real world. Furthermore, the AI community should not assume that agents will be naturally willing to cooperate with humans, or that cooperative behaviors in any one scenario necessarily translates to generalized willingness to cooperate in many other contexts.

To that end, our experimental setups (especially in Study 1 and Study 2) establish a robust methodology to examine human-machine teamwork by testing co-operative dynamics and coordination in multi-agent systems. Leveraging game theory in this way is useful precisely because deviations from Nash Equilibria highlight the distinctive dynamics governing cooperation and coordination in multi-agent systems, which in this context can serve as useful proxies for human-machine teamwork. This approach is very valuable for AI-safety researchers because it enables the precise and quantitative analysis of the cooperative dynamics between humans and AIs.

Overall, this dissertation extends prior findings of the value of game theory to study the behavior of multi-agent systems by showing how game theory can inform the design of experimental setups to identify the different ways humans and AIs can cooperate and coordinate under different incentive structures. Specifically, our methodology provides a strong and reliable basis for future researchers to test and make inferences about the ways in which different RL models, different beliefs about whether the other player is an AI or a human, as well as different game theory models with different Nash Equilibria influence the cooperative dynamics of human-machine teams.

### 10.3.2 Team cognition

This dissertation strongly emphasises the need to approach human-machine teamwork not from just the human or only the machine perspective, but from both. Human-machine teamwork is a complex phenomenon that, just as the studies show, varies wildly in different contexts and at different scales. Thus, this dissertation shows that human-machine teamwork can only be understood through an interdisciplinary perspective that includes both cognitive science as well as computer science.

Specifically, our results go contrary to the expectations traditionally set by the teamwork literature. For instance, Study 1, 2 and 3 invalidated the anticipated need for advanced natural language processing to effectively study human-machine teamwork, thereby showing that human-machine teamwork can emerge despite lack of communication [Bates and Weischedel, 2006]. Specifically, this dissertation showed how RL and our methodology for developing agents obviate the need for natural language processing because human-machine teamwork between humans and RL agents stems not from communication, as is the case with humans, but from cooperation and coordination that is much more akin to the game theoretical dynamics in multi-agent systems.

Besides communication, however, human-machine teams should make us reconsider what we know about situational awareness and its sources. Fan et al's (2010) speculated that situational awareness in human-machine teams would come with higher cognitive load; our data does not support this claim, as situational awareness did not result in fatigue and lower levels of team satisfaction as measured by the survey. Similarly, our data on situational awareness as measured by NeoCITIES does not highlight the coordination deficiencies resulting from a lack of direct communication in human-machine teams as suggested by Demir et al (2017); it shows precisely the opposite insofar as the human-machine teams displayed higher levels of situational awareness than human-only teams where communication was more frequent. Indeed, the superior performance exhibited by the machine-only teams in NeoCITIES shows that sequencing and synchronicity improve alongside the RL agent's skill-level in the simulation.

More importantly, this dissertation underscores the difference between automation and autonomy as outlined in Chapter 2. The prior literature makes several assumptions with regards to the alleged pivotal role predictability and directability play in human-machine teamwork [Klien et al., 2004, Christoffersen and Woods, 2002]. Such assumptions tend to be predicated upon an automation-centric view of AI and human-machine teamwork because it presupposes that the best way to achieve human-machine teamwork is by designing agents entirely around human dynamics and behaviors. Instead, this dissertation framed the human-machine team as a human-autonomy problem, thereby sidestepping many of these design assumptions in order to incorporate cutting-edge AI architecture without limiting researchers to any particular model. The subsequent results ended up aligning with McNeese et al's (2019) theory that humans and machines can retain interdependence and yet still function effectively as long as they share a common goal.

### 10.3.3 Collective intelligence

This dissertation strongly advocates for the design of prediction markets with human factors principles. Prediction markets are collective intelligence mechanisms, and thus only a stronger foundation in cognitive science can inform the design of better prediction markets by identifying the underlying mechanisms behind collective intelligence. However, this dissertation also shows that researchers ought to also approach prediction markets from a multi-agent teamwork perspective in order to discover better ways to accommodate the introduction of artificial intelligence so that humans and machines can work together and thus produce higher levels of collective intelligence.

Specifically, our results suggest that we have only scratched the surface of the potential designs for trading bots in prediction markets. Specifically, the prior literature has assumed that it should be up to the human participants to place bets to realign prices with historical base rates [Atanasov et al., 2016]. For example, a human-only prediction market expects that whenever the prices of “yes” and “no” shares don’t add up to 1, then human traders would buy and sell those shares until the prices converge towards the right values. This assumption however stands in sharp contrast with the marginal trader hypothesis, for the capital accumulated by essentially arbitraging temporary inconsistencies between the prices of “yes” and “no” shares would give those traders more resources to influence future prices despite the lack of better information. It thus makes sense to think have bots perform that function, thereby removing noise from a participant’s behavior as they therefore become limited in making directional bets as opposed to speculative trades.

Many other possibilities exist that can inspire future research, such as mean-reversion bots who buy and sell shares as they deviate from a moving average. More specifically however, the design of trading bots aimed at enhancing the collective intelligence of a prediction market should be informed by the biases the literature has identified in prediction markets. Our design was very much informed by Dudik et al’s (2017) in classifying the sources and impacts of different biases. Since sampling error arises from the noisiness of the local knowledge of the participants, we included the randomized trading bots so that the direction of the variation would not be biased in any particular direction. Since market-maker bias arises from players overshooting or undershooting their estimates based on how they are rewarded by functions used to facilitate trading, we relied upon the volatility emerging from the bots trading to incentive initial trading, which would not exhibit any specific pattern detectable and exploitable by the human participants. Lastly, since converge error arises from the market price constantly fluctuating, we included the other type of trading bot whose entire objective is to stabilize prices by removing excess volatility from the system.

## 10.4 Future Work

As part of this dissertation, several platforms were developed in order to run the experiments. Each platform focused on a particular dimension of human-machine teamwork: the Game Theory for Teams environment used in Study 1 and Study 2 enables the study of human-machine teamwork through game theoretical scenarios and multi-agent systems, the updated version of NeoCITIES used in Study 3 enables human-machine teams to be studied through a simulation that has been validated by the prior literature as a test-bed for team cognition, and the prediction market used in Study 4 provides a more effective way to study collective intelligence. Each platform opens up many exciting opportunities for future research

### 10.4.1 Game Theory for Teams Environment

For future work related to human-machine teamwork in multi-agent systems, it would be useful to extend the number of game theoretical scenarios available to researchers. Specifically, Study 1 and Study 2 showed how it's possible to isolate and separately study coordination and cooperation through game theory models whose incentive structure emphasis a particular dynamic. Similarly, other game theory models exist that can be useful for future researchers to identify specific multi-agent system dynamics.

For example, an interesting game that can be used to advance human-machine teamwork research is the diner's dilemma, which posits a scenario where the players have decided to split the cost equally and are confronted with the choice to either order a more expensive meal or a cheap alternative. The central conflict is that if all players pick the more expensive meal, then all the members of the group experience a loss caused by the higher bill paid by each individual, yet an individual player can experience a gain by successfully ordering the expensive meal when everyone else orders the cheap alternative. The game's structure supports an infinite number of players, thereby providing researchers with the opportunity to test how cooperation is affected as the human-machine team grows larger.

Under the same theme, the public goods game would also provide interesting research opportunities. Often used in experimental economics, the game involves players depositing private tokens into a public pot that is then multiplied by a factor and subsequently evenly distributed among all players, regardless of participation. The central conflict of the game is the free-rider problem, which directly deals with trust. Therefore, researchers could implement the public goods game in the environment and test how sensitive trust in a human-machine team is to various factors, including the factor by which the common pot is multiplied.

### 10.4.2 NeoCITIES

With respect to NeoCITIES, there are several modifications that can expand the simulation's capabilities in order to support a wider variety of experiments. At the outset, Study 3 was limited to a fairly basic scenario found in the literature, but the platform supports far more complex scenarios with more resources, different priority levels, and stricter roles. Beyond scenarios, NeoCITIES currently supports a theoretically unlimited number of players, although the RL agents need to be trained differently for such scenarios, especially if they involve more sharply differentiated roles than those featured in Study 3.

Separately, future work may involve limiting communication among all players, in order to equalize the playing field between team types. Specifically, the human participant in a human-machine-machine team is not able to communicate with their teammates, while communication is possible between the two human participants in a human-human-machine team. Given the prominent role communication plays in human coordination, new results may emerge when humans are forced to rely on indirect means of coordination irrespective of team type.

Lastly, future work should focus on breaking down team scores and other measurements by player, in order to better identify the relative value each teammate brings to the team. Indeed, this kind of in depth analysis should motivate the testing of a wider variety of RL models across more differentiated roles than in Study 3. Specifically, individualizing the team scores may help shed light as to whether the relationship between teammates is even (they each contribute the same amount to the team score) or uneven (some teammates contribute a disproportionate amount to the team score). The relationship between contribution inequality and team performance, especially if mediated by team types, would produce valuable insights into the optimal equilibrium state in human-machine teams.

### 10.4.3 Prediction Market

The prediction market developed for Study 4 enables a more precise study of collective intelligence by providing both the ground truth of events being forecasted as well as the local knowledge available to each participant. These two mechanics enable several promising avenues for future research.

To start with, a follow up to Study 4 would be to simulate knowledge updates among the participants to compare human-only and hybrid prediction markets in terms of market efficiency as measured by how quickly event prices update to incorporate changes in the local knowledge of the participants. Furthermore, the prediction market equivalent of bubbles can be simulated by distorting the local knowledge allocation in order to dilute the signal possessed by the participants. The more players trade on information that is a duplicate of other players' information, the more the event price will diverge from ground truth until an update causes it to crash. Such scenarios of low probability but high magnitude risk have never been able to be studied for the purposes of collective intelligence and forecasting, and thus provide fruitful ground for future research.

Separately, another promising area for future research would be demographics. Given that our prediction market can map onto real world events just as it can on simulated ones, researchers could implement an instance of the prediction market to study forecasting among particular demographics to determine the ways in which collective intelligence is influenced by social context. For example, future research may delve into the differences between a prediction market where only doctors participate to one where only nurses participate as they both try to forecast the influx of patients at a hospital. Furthermore, such research would also include an analysis of the impact of varying demographic distributions among the traders would have on the AI's ability to effectively aggregate information to calibrate the prediction market's forecasts.

## 10.5 Closing Remarks

The research objectives for this dissertation were wide-ranging. Specifically, this dissertation sought to integrate multi-agent systems, team cognition, and collective intelligence through an interdisciplinary approach directed at advancing each of the research communities. looking back upon the work within this dissertation, the objectives were effectively addressed.

The first goal of the dissertation was the identification of the similarities and differences between human-machine teamwork, human-human teamwork, and machine-machine teamwork. The dissertation shows that human-machine teams behave more closely to machine-machine teams as opposed to human teams, and thus often coordinate and outperform despite the inability to directly communicate.

The second goal was to explore the unique ways in which team cognition emerges in human-machine teams. Contrary to what was anticipated, team cognition did not prove to be an effective model to predict human-machine team performance with. These results run counter to what would be expected from the team cognition literature.

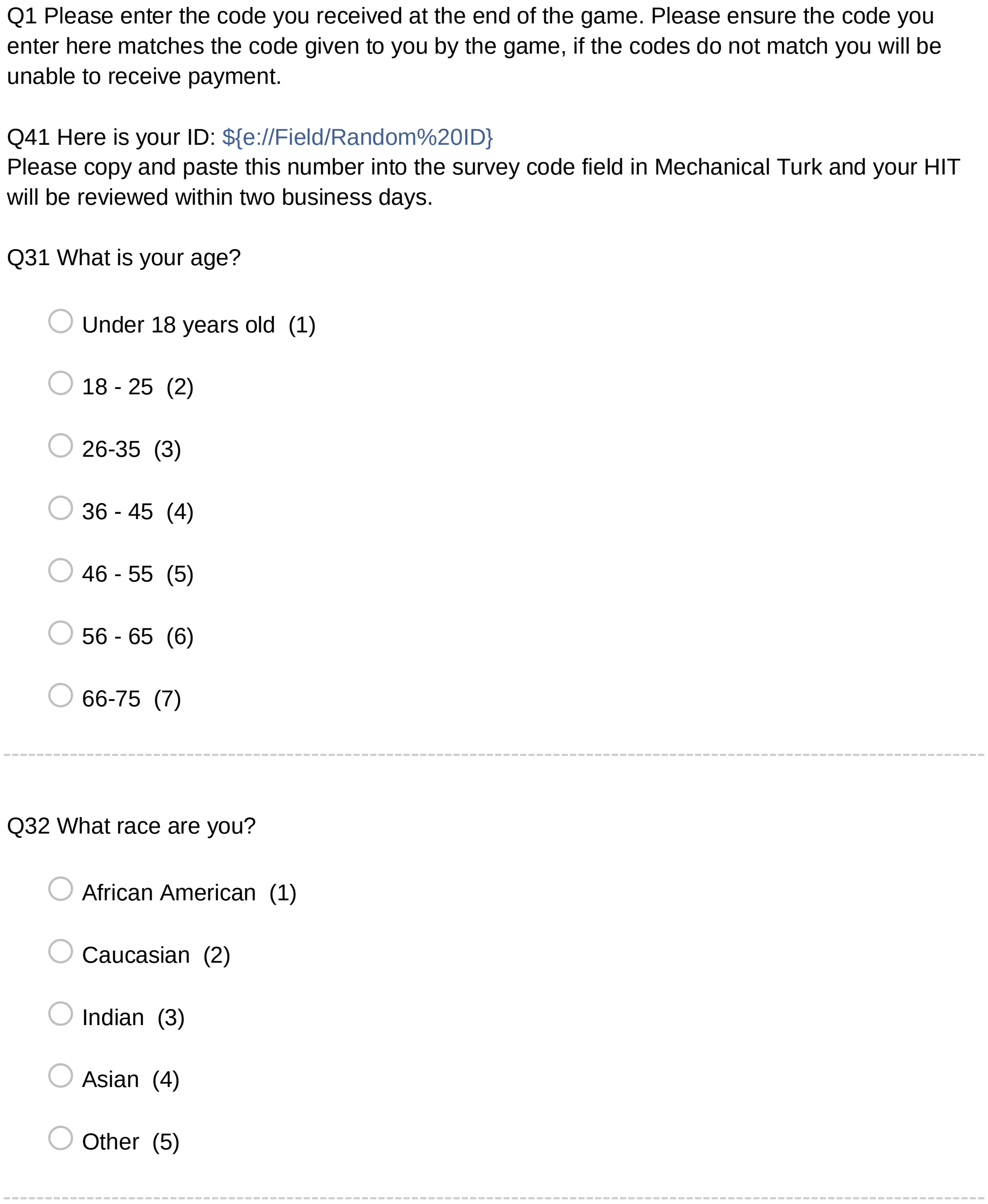
The third goal was to identify the ways in which AI can enhance collective intelligence in human-machine teams. This dissertation shows that AI can play the dual role of participant and aggregate in a prediction market, and that doing so meaningfully enhances collective intelligence as measured by the prediction market's heightened ability to forecast despite uncertainty.

The last goal was to develop empirically-backed design guidelines for the integration of AI in prediction markets. This dissertation successfully implemented Model A, which used human factors principle to design a prediction market that addresses behavioral biases while simultaneously preserving the key features that give rise to collective intelligence. The major takeaway is that by thinking of a prediction market as a higher-level form of human-machine teamwork, it becomes possible to discover new ways for humans and machines to work together to tackle a problem as challenging as probabilistic decision-making.

Overall, in the process of achieving these goals, the dissertation produced three sophisticated research platforms that will open new ways for future researchers to study multi-agent systems, human-machine teamwork, and collective intelligence in prediction markets.

# Appendices

# Appendix A Team Cognition Survey



Q1 Please enter the code you received at the end of the game. Please ensure the code you enter here matches the code given to you by the game, if the codes do not match you will be unable to receive payment.

Q41 Here is your ID: ${e://Field/Random%20ID}
Please copy and paste this number into the survey code field in Mechanical Turk and your HIT will be reviewed within two business days.

Q31 What is your age?

○ Under 18 years old (1)

○ 18 - 25 (2)

○ 26-35 (3)

○ 36 - 45 (4)

○ 46 - 55 (5)

○ 56 - 65 (6)

○ 66-75 (7)

Q32 What race are you?

○ African American (1)

○ Caucasian (2)

○ Indian (3)

○ Asian (4)

○ Other (5)

Q33 What is your gender?

○ Male (1)

○ Female (2)

○ Other (3)

Q39 Type the response "human".

________________________________________________________________

Q34 What is the highest degree or level of school you have completed? (If you're currently enrolled in school, please state the highest degree you have *received)*

○ Less than a high school diploma (1)

○ High school diploma (GED) (2)

○ Some college (3)

○ Associates degree (AA) (4)

○ Bachelor's degree (BA, BS) (5)

○ Masters degree (MA, MS, MEd) (6)

○ Professional degree (MD, DDS, DVM) (7)

○ Doctorate degree (PhD, EdD) (8)

Q2 How mentally demanding was the task?

○ Very Low 1 (1)

○ Below Average 2 (2)

○ Average 3 (3)

○ Above Average 4 (4)

○ Very High 5 (5)

Q36 Select the choice that is second to the furthest right.

○ Very Low 1 (6)

○ Below Average 2 (7)

○ Average 3 (8)

○ Above Average 4 (9)

○ Very High 5 (10)

Q3 How physically demanding was the task?

○ Very Low 1 (1)

○ Below Average 2 (2)

○ Average 3 (3)

○ Above Average 4 (4)

○ Very High 5 (5)

Q4 How hurried or rushed was the pace of the task?

○ Very Low 1 (1)

○ Below Average 2 (2)

○ Average 3 (3)

○ Above Average 4 (4)

○ Very High 5 (5)

Q5 How successful were you in accomplishing what you were asked to do?

○ Very Low 1 (1)

○ Below Average 2 (2)

○ Average 3 (3)

○ Above Average 4 (4)

○ Very High 5 (5)

Q35 Select the choice that is all the way left.

○ Very Low 1 (6)

○ Below Average 2 (7)

○ Average 3 (8)

○ Above Average 4 (9)

○ Very High 5 (10)

Q6 How hard did you have to work to accomplish your level of performance?

○ Very Low 1 (1)

○ Below Average 2 (2)

○ Average 3 (3)

○ Above Average 4 (4)

○ Very High 5 (5)

Q7 How insecure, discouraged, irritated, stressed, and annoyed were you?

○ Very Low 1 (1)

○ Below Average 2 (2)

○ Average 3 (3)

○ Above Average 4 (4)

○ Very High 5 (5)

Q42 My team understands its roles and responsibilities.

○ Very Low (1)

○ Below Average (2)

○ Average (3)

○ Above Average (4)

○ Very High (5)

Q43 My team knows where it can get information

○ Very Low (1)

○ Below Average (2)

○ Average (3)

○ Above Average (4)

○ Very High (5)

Q44 My team understands interaction patterns

○ Very Low (1)

○ Below Average (2)

○ Average (3)

○ Above Average (4)

○ Very High (5)

Q45 My team understands how they can exchange information for doing various team tasks

○ Very Low (1)

○ Below Average (2)

○ Average (3)

○ Above Average (4)

○ Very High (5)

Q46 My team can adopt flexibly to any roles within the team

○ Very Low (1)

○ Below Average (2)

○ Average (3)

○ Above Average (4)

○ Very High (5)

Q47 My team is likely to make a decision together

○ Very Low (1)

○ Below Average (2)

○ Average (3)

○ Above Average (4)

○ Very High (5)

Q48 My team communicates with other teammates while performing team tasks

○ Very Low (1)

○ Below Average (2)

○ Average (3)

○ Above Average (4)

○ Very High (5)

Q49 My teammates informally communicate with one another throughout various team tasks.

○ Very Low (1)

○ Below Average (2)

○ Average (3)

○ Above Average (11)

○ Very High (12)

Q50 My teammates consistently demonstrate effective listening skills.

○ Very Low (1)

○ Below Average (2)

○ Average (3)

○ Above Average (4)

○ Very High (5)

Q51 My teammate has a general knowledge of specific team tasks.

○ Very Low (1)

○ Below Average (2)

○ Average (3)

○ Above Average (4)

○ Very High (5)

Q52 My teammate knows specific strategies for completing various tasks

○ Very Low (1)

○ Below Average (2)

○ Average (3)

○ Above Average (4)

○ Very High (5)

Q53 My teammate knows the general process involved in conducting a given task

○ Very Low (1)

○ Below Average (2)

○ Average (3)

○ Above Average (4)

○ Very High (5)

Q54 My teammate understands the skills necessary for doing various team tasks.

○ Very Low (1)

○ Below Average (2)

○ Average (3)

○ Above Average (4)

○ Very High (5)

Q55 My teammate communicates with other teammates while performing team tasks

○ Very Low (1)

○ Below Average (2)

○ Average (3)

○ Above Average (4)

○ Very High (5)

Q56 My teammate supports continuous improvement in terms of personal skills as well as overall team skills

○ Very Low (1)

○ Below Average (2)

○ Average (3)

○ Above Average (4)

○ Very High (5)

Q57 I have a good knowledge about my teammates' attitudes

○ Very Low (1)

○ Below Average (2)

○ Average (3)

○ Above Average (4)

○ Very High (5)

Q58 There is a sense of cohesion and cooperation among my teammates

○ Very Low (1)

○ Below Average (2)

○ Average (3)

○ Above Average (4)

○ Very High (5)

Q59 My teammate takes pride in his/her work

○ Very Low (1)

○ Below Average (2)

○ Average (3)

○ Above Average (4)

○ Very High (5)

Q60 I have a good knowledge about my teammates’ preferences

○ Very Low (1)

○ Below Average (2)

○ Average (3)

○ Above Average (4)

○ Very High (5)

Q61 My teammates like to do various team tasks

○ Very Low (1)

○ Below Average (2)

○ Average (3)

○ Above Average (4)

○ Very High (5)

Q62 My teammates enjoy thinking

○ Very Low (1)

○ Below Average (2)

○ Average (3)

○ Above Average (4)

○ Very High (5)

Q63 I have a good knowledge about my teammates' tendencies

○ Very Low (1)

○ Below Average (2)

○ Average (3)

○ Above Average (4)

○ Very High (5)

---

Q64 My teammates are committed to the team goal

○ Very Low (1)

○ Below Average (2)

○ Average (3)

○ Above Average (4)

○ Very High (5)

---

Q65 My teammates encourage each other's work to improve various team task outcomes

○ Very Low (1)

○ Below Average (2)

○ Average (3)

○ Above Average (4)

○ Very High (5)

---

Q66 My teammate strives to express his or her opinion

○ Very Low (1)

○ Below Average (2)

○ Average (3)

○ Above Average (4)

○ Very High (5)

Q67 There is an atmosphere of trust among my teammates

○ Very Low (1)

○ Below Average (2)

○ Average (3)

○ Above Average (4)

○ Very High (5)

Q68 My team creates a work environment that promotes productive results

○ Very Low (1)

○ Below Average (2)

○ Average (3)

○ Above Average (4)

○ Very High (5)

Q69 My team creates a safe environment to openly discuss any issue related to the team's success

○ Very Low (1)

○ Below Average (2)

○ Average (3)

○ Above Average (4)

○ Very High (5)

Q70 My team acknowledges and rewards behaviors that contribute to an open team climate

○ Very Low (1)

○ Below Average (2)

○ Average (3)

○ Above Average (4)

○ Very High (5)

Q71 My team often utilizes different opinions for the sake of obtaining optimal outcomes

○ Very Low (1)

○ Below Average (2)

○ Average (3)

○ Above Average (4)

○ Very High (5)

Q72 Discussions for decision making occur within my team during meetings so that team meetings are viewed as useful

○ Very Low (1)

○ Below Average (2)

○ Average (3)

○ Above Average (4)

○ Very High (5)

Q74 My team knows the environmental constraints when we perform various team tasks

○ Very Low (1)

○ Below Average (2)

○ Average (3)

○ Above Average (4)

○ Very High (5)

Q12 Please state your agreement to the following statements as they relate to yourself.

Q9 I am optimistic towards artificial intelligence.

○ Strongly disagree (1)

○ Somewhat disagree (2)

○ Neither agree nor disagree (3)

○ Somewhat agree (4)

○ Strongly agree (5)

Q10 I am experienced working with artificial intelligence.

○ Strongly disagree (1)

○ Somewhat disagree (2)

○ Neither agree nor disagree (3)

○ Somewhat agree (4)

○ Strongly agree (5)

Q11 I have a positive relationship with technology.

○ Strongly disagree (1)

○ Somewhat disagree (2)

○ Neither agree nor disagree (3)

○ Somewhat agree (4)

○ Strongly agree (5)

Q37 Select the middle answer choice.

- ○ Strongly disagree (27)
- ○ Somewhat disagree (28)
- ○ Neither agree nor disagree (29)
- ○ Somewhat agree (30)
- ○ Strongly agree (31)

Q13 I would prefer working with a human rather than an AI.

- ○ Strongly disagree (1)
- ○ Somewhat disagree (2)
- ○ Neither agree nor disagree (3)
- ○ Somewhat agree (4)
- ○ Strongly agree (5)

Q15 Please rate your agreement to the following statements as they pertain to yourself.

Q16 I felt my teammate and I had a shared understanding of our teamwork.

- ○ Strongly disagree (1)
- ○ Somewhat disagree (2)
- ○ Neither agree nor disagree (3)
- ○ Somewhat agree (4)
- ○ Strongly agree (5)

Q17 I tried to cooperate with my teammate during the task.

- ○ Strongly disagree (1)
- ○ Somewhat disagree (2)
- ○ Neither agree nor disagree (3)
- ○ Somewhat agree (4)
- ○ Strongly agree (5)

Q18 My team member was cooperative during the task.

- ○ Strongly disagree (1)
- ○ Somewhat disagree (2)
- ○ Neither agree nor disagree (3)
- ○ Somewhat agree (4)
- ○ Strongly agree (5)

Q19 My teammate and I were cooperative during the task.

- ○ Strongly disagree (1)
- ○ Somewhat disagree (2)
- ○ Neither agree nor disagree (3)
- ○ Somewhat agree (4)
- ○ Strongly agree (5)

Q38 You must select the answer choice that states somewhat agree.

○ Strongly disagree (27)

○ Somewhat disagree (28)

○ Neither agree nor disagree (29)

○ Somewhat agree (30)

○ Strongly agree (31)

---

Q20 My teammate and I worked effectively during the task.

○ Strongly disagree (1)

○ Somewhat disagree (2)

○ Neither agree nor disagree (3)

○ Somewhat agree (4)

○ Strongly agree (5)

---

Q21 My teammate and I worked together better at the end of the task than at the beginning.

○ Strongly disagree (1)

○ Somewhat disagree (2)

○ Neither agree nor disagree (3)

○ Somewhat agree (4)

○ Strongly agree (5)

---

Q22 I am satisfied with my performance.

○ Strongly disagree  (1)

○ Somewhat disagree  (2)

○ Neither agree nor disagree  (3)

○ Somewhat agree  (4)

○ Strongly agree  (5)

Q23 I am satisfied with my teammate's performance.

○ Strongly disagree  (1)

○ Somewhat disagree  (2)

○ Neither agree nor disagree  (3)

○ Somewhat agree  (4)

○ Strongly agree  (5)

Q24 I would work with my teammate on another task.

○ Strongly disagree  (1)

○ Somewhat disagree  (2)

○ Neither agree nor disagree  (3)

○ Somewhat agree  (4)

○ Strongly agree  (5)

Q25 I felt I was the leader during our collaboration.

○ Strongly disagree (1)

○ Somewhat disagree (2)

○ Neither agree nor disagree (3)

○ Somewhat agree (4)

○ Strongly agree (5)

Q26 I trust my teammate in acting in our mutual best interest.

○ Strongly disagree (1)

○ Somewhat disagree (2)

○ Neither agree nor disagree (3)

○ Somewhat agree (4)

○ Strongly agree (5)

Q27 I reacted aggressively towards my teammate.

○ Strongly disagree (1)

○ Somewhat disagree (2)

○ Neither agree nor disagree (3)

○ Somewhat agree (4)

○ Strongly agree (5)

Q28 My teammate reacted aggressively towards me.

○ Strongly disagree  (1)

○ Somewhat disagree  (2)

○ Neither agree nor disagree  (3)

○ Somewhat agree  (4)

○ Strongly agree  (5)

Q29 I would like to be able to communicate with my teammate.

○ Strongly disagree  (1)

○ Somewhat disagree  (2)

○ Neither agree nor disagree  (3)

○ Somewhat agree  (4)

○ Strongly agree  (5)

Q30 Our performance would benefit from?

________________________________________________________________

# Bibliography

[Allen and Ferguson, 2002] Allen, J. and Ferguson, G. (2002). Human-machine collaborative planning. In *Proceedings of the Third International NASA Workshop on Planning and Scheduling for Space*, pages 27–29.

[Atanasov et al., 2013] Atanasov, P., Rescober, P., Stone, E., Servan-Schreiber, E., Mellers, B., Tetlock, P., and Ungar, L. (2013). The marketcast method for aggregating prediction market forecasts. In *International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction*, pages 28–37. Springer.

[Atanasov et al., 2016] Atanasov, P., Rescober, P., Stone, E., Swift, S. A., Servan-Schreiber, E., Tetlock, P., Ungar, L., and Mellers, B. (2016). Distilling the wisdom of crowds: Prediction markets vs. prediction polls. *Management science*, 63(3):691–706.

[Attarwala et al., 2017] Attarwala, A., Dimitrov, S., and Obeidi, A. (2017). How efficient is twitter: Predicting 2012 us presidential elections using support vector machine via twitter and comparing against iowa electronic markets. In *2017 Intelligent Systems Conference (IntelliSys)*, pages 646–652. IEEE.

[Bab and Brafman, 2008] Bab, A. and Brafman, R. I. (2008). Multi-agent reinforcement learning in common interest and fixed sum stochastic games: An experimental study. *Journal of Machine Learning Research*, 9(Dec):2635–2675.

[Ball et al., 2010] Ball, J., Myers, C., Heiberg, A., Cooke, N. J., Matessa, M., Freiman, M., and Rodgers, S. (2010). The synthetic teammate project. *Computational and Mathematical Organization Theory*, 16(3):271–299.

[Bansal et al., 2017] Bansal, T., Pachocki, J., Sidor, S., Sutskever, I., and Mordatch, I. (2017). Emergent complexity via multi-agent competition. *arXiv preprint arXiv:1710.03748*.

[Barberis Canonico et al., 2019a] Barberis Canonico, L., McNeese, N. J., and Flathmann, C. (2019a). Collectively intelligent teams: Integrating team cognition, collective intelligence, and ai for future teaming. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 63. SAGE Publications Sage CA: Los Angeles, CA.

[Barberis Canonico et al., 2019b] Barberis Canonico, L., McNeese, N. J., and Flathmann, C. (2019b). Human-ai teams as multi-agent systems: Human-machine teamwork through game theory. *Journal of Cognitive Engineering and Decision Making.*

[Barberis Canonico et al., 2019c] Barberis Canonico, L., McNeese, N. J., and Flathmann, C. (2019c). The wisdom of the market: Using human factors to design prediction markets for collective intelligence. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 63. SAGE Publications Sage CA: Los Angeles, CA.

[Barberis Canonico et al., 2019d] Barberis Canonico, L., McNeese, N. J., Flathmann, C., and Schelble, B. (2019d). Game theory for teams: Using game theory models to understand human-ai teamwork. *IEEE Transactions on Human-Machine Systems.*

[Barbu and Lay, 2012] Barbu, A. and Lay, N. (2012). An introduction to artificial prediction markets for classification. *Journal of Machine Learning Research*, 13(Jul):2177–2204.

[Bates et al., 2014] Bates, D., Mächler, M., Bolker, B., and Walker, S. (2014). Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823.*

[Bates and Weischedel, 2006] Bates, M. and Weischedel, R. M. (2006). *Challenges in natural language processing.* Cambridge University Press.

[Bozdog et al., 2011] Bozdog, D., Florescu, I., Khashanah, K., and Wang, J. (2011). Rare events analysis for high-frequency equity data. *Wilmott*, 2011(54):74–81.

[Cannon-Bowers et al., 1990] Cannon-Bowers, J. A., Salas, E., and Converse, S. (1990). Cognitive psychology and team training: Training shared mental models and complex systems. *Human factors society bulletin*, 33(12):1–4.

[Chattopadhyay et al., 2017] Chattopadhyay, P., Yadav, D., Prabhu, V., Chandrasekaran, A., Das, A., Lee, S., Batra, D., and Parikh, D. (2017). Evaluating visual conversational agents via cooperative human-ai games. In *Fifth AAAI Conference on Human Computation and Crowdsourcing.*

[Chen et al., 2005] Chen, Y., Chu, C.-H., Mullen, T., and Pennock, D. M. (2005). Information markets vs. opinion pools: An empirical comparison. In *Proceedings of the 6th ACM conference on Electronic commerce*, pages 58–67. ACM.

[Chen et al., 2010] Chen, Y., Dimitrov, S., Sami, R., Reeves, D. M., Pennock, D. M., Hanson, R. D., Fortnow, L., and Gonen, R. (2010). Gaming prediction markets: Equilibrium strategies with a market maker. *Algorithmica*, 58(4):930–969.

[Chen and Vaughan, 2010] Chen, Y. and Vaughan, J. W. (2010). A new understanding of prediction markets via no-regret learning. In *Proceedings of the 11th ACM conference on Electronic commerce*, pages 189–198. ACM.

[Christoffersen and Woods, 2002] Christoffersen, K. and Woods, D. D. (2002). How to make automated systems team players. In *Advances in human performance and cognitive engineering research*, pages 1–12. Emerald Group Publishing Limited.

[Cooke et al., 2012] Cooke, N. J., Gorman, J. C., Myers, C., and Duran, J. (2012). Theoretical underpinnings of interactive team cognition. *Theories of team cognition: Cross-disciplinary perspectives*, pages 187–207.

[Cooke et al., 2007] Cooke, N. J., Gorman, J. C., Winner, J. L., and Durso, F. (2007). Team cognition. *Handbook of applied cognition*, 2:239–268.

[Cords, 2007] Cords, S. S. (2007). How life imitates chess: Making the right moves-from the board to the boardroom.

[Crowder and Carbone, 2014] Crowder, J. A. and Carbone, J. N. (2014). Collaborative shared awareness: human-ai collaboration. In *Proceedings of the International Conference on Information and Knowledge Engineering (IKE)*, page 1. The Steering Committee of The World Congress in Computer Science, Computer . . . .

[de Cote et al., 2006] de Cote, E. M., Lazaric, A., and Restelli, M. (2006). Learning to cooperate in multi-agent social dilemmas. In *AAMAS*, volume 6, pages 783–785.

[Demir et al., 2016] Demir, M., McNeese, N. J., and Cooke, N. J. (2016). Team communication behaviors of the human-automation teaming. In *2016 IEEE International Multi-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support (CogSIMA)*, pages 28–34. IEEE.

[Demir et al., 2017] Demir, M., McNeese, N. J., and Cooke, N. J. (2017). Team situation awareness within the context of human-autonomy teaming. *Cognitive Systems Research*, 46:3–12.

[Demir et al., 2018] Demir, M., McNeese, N. J., and Cooke, N. J. (2018). The impact of perceived autonomous agents on dynamic team behaviors. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2(4):258–267.

[Dudik et al., 2017] Dudik, M., Lahaie, S., Rogers, R. M., and Vaughan, J. W. (2017). A decomposition of forecast error in prediction markets. In *Advances in Neural Information Processing Systems*, pages 4371–4380.

[Duhigg, 2016] Duhigg, C. (2016). What google learned from its quest to build the perfect team. *The New York Times Magazine*, 26:2016.

[Edwards et al., 2006] Edwards, B. D., Day, E. A., Arthur Jr, W., and Bell, S. T. (2006). Relationships among team ability composition, team mental models, and team performance. *Journal of Applied Psychology*, 91(3):727.

[Endsley, 2015] Endsley, M. R. (2015). Autonomous horizons: System autonomy in the air force-a path to the future. *United States Air Force Office of the Chief Scientist, AF/ST TR*, pages 15–01.

[Erev and Roth, 1998] Erev, I. and Roth, A. E. (1998). Predicting how people play games: Reinforcement learning in experimental games with unique, mixed strategy equilibria. *American economic review*, pages 848–881.

[Eriksson and Simpson, 2010] Eriksson, K. and Simpson, B. (2010). Emotional reactions to losing explain gender differences in entering a risky lottery. *Judgment and Decision Making*, 5(3):159.

[Fiore et al., 2010] Fiore, S. M., Rosen, M. A., Smith-Jentsch, K. A., Salas, E., Letsky, M., and Warner, N. (2010). Toward an understanding of macrocognition in teams: Predicting processes in complex collaborative contexts. *Human Factors*, 52(2):203–224.

[Foerster et al., 2018] Foerster, J., Chen, R. Y., Al-Shedivat, M., Whiteson, S., Abbeel, P., and Mordatch, I. (2018). Learning with opponent-learning awareness. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, pages 122–130. International Foundation for Autonomous Agents and Multiagent Systems.

[Gao et al., 2016] Gao, F., Cummings, M., and Solovey, E. (2016). Designing for robust and effective teamwork in human-agent teams. In *Robust intelligence and trust in autonomous systems*, pages 167–190. Springer.

[Goel et al., 2010] Goel, S., Reeves, D. M., Watts, D. J., and Pennock, D. M. (2010). Prediction without markets. In *Proceedings of the 11th ACM conference on Electronic commerce*, pages 357–366. ACM.

[Goertzel and Pennachin, 2007] Goertzel, B. and Pennachin, C. (2007). *Artificial general intelligence*, volume 2. Springer.

[Gosling et al., 2004] Gosling, S. D., Vazire, S., Srivastava, S., and John, O. P. (2004). Should we trust web-based studies? a comparative analysis of six preconceptions about internet questionnaires. *American psychologist*, 59(2):93.

[Graham et al., 2004] Graham, J., Schneider, M., Bauer, A., Bessiere, K., and Gonzalez, C. (2004). Shared mental models in military command and control organizations: Effect of social network distance. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 48, pages 509–512. SAGE Publications Sage CA: Los Angeles, CA.

[Grocer, 2010] Grocer, S. (2010). Senators seek regulators' report on causes of market volatility. *WallStreet Journal, May*, 7:5.

[Gu et al., 2016] Gu, S., Lillicrap, T., Sutskever, I., and Levine, S. (2016). Continuous deep q-learning with model-based acceleration. In *International Conference on Machine Learning*, pages 2829–2838.

[Hamilton et al., 2010] Hamilton, K., Mancuso, V., Minotra, D., Hoult, R., Mohammed, S., Parr, A., Dubey, G., McMillan, E., and McNeese, M. (2010). Using the neocities 3.1 simulation to study and measure team cognition. In *Proceedings of the human factors and ergonomics society annual meeting*, volume 54, pages 433–437. SAGE Publications Sage CA: Los Angeles, CA.

[Hancock and Szalma, 2008] Hancock, P. A. and Szalma, J. L. (2008). *Performance under stress*. Ashgate Publishing, Ltd.

[Hanson, 2003] Hanson, R. (2003). Combinatorial information market design. *Information Systems Frontiers*, 5(1):107–119.

[Hellar and McNeese, 2010] Hellar, D. B. and McNeese, M. (2010). Neocities: A simulated command and control task environment for experimental research. In *Proceedings of the human factors and ergonomics society annual meeting*, volume 54, pages 1027–1031. SAGE Publications Sage CA: Los Angeles, CA.

[Hester et al., 2018] Hester, T., Vecerik, M., Pietquin, O., Lanctot, M., Schaul, T., Piot, B., Horgan, D., Quan, J., Sendonaris, A., Osband, I., et al. (2018). Deep q-learning from demonstrations. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[Hinsz et al., 1997] Hinsz, V. B., Tindale, R. S., and Vollrath, D. A. (1997). The emerging conceptualization of groups as information processors. *Psychological bulletin*, 121(1):43.

[Hipp et al., 2011] Hipp, J., Flotte, T., Monaco, J., Cheng, J., Madabhushi, A., Yagi, Y., Rodriguez-Canales, J., Emmert-Buck, M., Dugan, M. C., Hewitt, S., et al. (2011). Computer aided diagnostic tools aim to empower rather than replace pathologists: Lessons learned from computational chess. *Journal of pathology informatics*, 2.

[Hollenbeck et al., 1998] Hollenbeck, J. R., Ilgen, D. R., LePine, J. A., Colquitt, J. A., and Hedlund, J. (1998). Extending the multilevel theory of team decision making: Effects of feedback and experience in hierarchical teams. *Academy of Management Journal*, 41(3):269–282.

[Horton et al., 2011] Horton, J. J., Rand, D. G., and Zeckhauser, R. J. (2011). The online laboratory: Conducting experiments in a real labor market. *Experimental economics*, 14(3):399–425.

[Hsu, 2004] Hsu, F.-H. (2004). *Behind Deep Blue: Building the computer that defeated the world chess champion.* Princeton University Press.

[Hu et al., 1998] Hu, J., Wellman, M. P., et al. (1998). Multiagent reinforcement learning: theoretical framework and an algorithm. In *ICML*, volume 98, pages 242–250. Citeseer.

[Ipeirotis, 2009] Ipeirotis, P. (2009). Turker demographics vs. internet demographics. *A Computer Scientist in a Business School*, 16.

[Ismail and Mnyanda, 2016] Ismail, N. I. and Mnyanda, L. (2016). Flash crash of the pound baffles traders with algorithms being blamed.

[Jahedpari et al., 2017] Jahedpari, F., Rahwan, T., Hashemi, S., Michalak, T. P., De Vos, M., Padget, J., and Woon, W. L. (2017). Online prediction via continuous artificial prediction markets. *IEEE Intelligent Systems*, 32(1):61–68.

[Kaur and Wasan, 2006] Kaur, H. and Wasan, S. K. (2006). Empirical study on applications of data mining techniques in healthcare. *Journal of Computer science*, 2(2):194–200.

[Kirilenko et al., 2017] Kirilenko, A., Kyle, A. S., Samadi, M., and Tuzun, T. (2017). The flash crash: High-frequency trading in an electronic market. *The Journal of Finance*, 72(3):967–998.

[Klien et al., 2004] Klien, G., Woods, D. D., Bradshaw, J. M., Hoffman, R. R., and Feltovich, P. J. (2004). Ten challenges for making automation a" team player" in joint human-agent activity. *IEEE Intelligent Systems*, 19(6):91–95.

[Klimoski and Mohammed, 1994] Klimoski, R. and Mohammed, S. (1994). Team mental model: Construct or metaphor? *Journal of management*, 20(2):403–437.

[Krantz and Dalal, 2000] Krantz, J. H. and Dalal, R. (2000). Validity of web-based psychological research. In *Psychological experiments on the Internet*, pages 35–60. Elsevier.

[Laskey et al., 2015] Laskey, K. B., Hanson, R., and Twardy, C. (2015). Combinatorial prediction markets for fusing information from distributed experts and models. In *2015 18th International Conference on Information Fusion (Fusion)*, pages 1892–1898. IEEE.

[Lee and Johnson, 2008] Lee, M. and Johnson, T. E. (2008). Understanding the effects of team cognition associated with complex engineering tasks: Dynamics of shared mental models, task-smm, and team-smm. *Performance Improvement Quarterly*, 21(3):73–95.

[Letsky et al., 2007] Letsky, M., Warner, N., Fiore, S. M., Rosen, M., and Salas, E. (2007). Macrocognition in complex team problem solving. Technical report, Office of Naval Research Arlington VA.

[Luckner et al., 2011] Luckner, S., Schröder, J., Slamka, C., Skiera, B., Spann, M., Weinhardt, C., Geyer-Schulz, A., and Franke, M. (2011). *Prediction markets: Fundamentals, designs, and applications*. Springer Science & Business Media.

[Malone, 2018] Malone, T. W. (2018). *Superminds: The surprising power of people and computers thinking together*. Little, Brown.

[Malone et al., 2009] Malone, T. W., Laubacher, R., and Dellarocas, C. (2009). Harnessing crowds: Mapping the genome of collective intelligence.

[Mathieu et al., 2000] Mathieu, J. E., Heffner, T. S., Goodwin, G. F., Salas, E., and Cannon-Bowers, J. A. (2000). The influence of shared mental models on team process and performance. *Journal of applied psychology*, 85(2):273.

[McComb et al., 2010] McComb, S., Kennedy, D., Perryman, R., Warner, N., and Letsky, M. (2010). Temporal patterns of mental model convergence: Implications for distributed teams interacting in electronic collaboration spaces. *Human Factors*, 52(2):264–281.

[McComb, 2007] McComb, S. A. (2007). Mental model convergence: The shift from being an individual to being a team member. In *Multi-level issues in organizations and time*, pages 95–147. Emerald Group Publishing Limited.

[McNeese et al., 2017] McNeese, M., McNeese, N. J., Endsley, T., Reep, J., and Forster, P. (2017). Simulating team cognition in complex systems: Practical considerations for researchers. In *Advances in Neuroergonomics and Cognitive Engineering*, pages 255–267. Springer.

[McNeese et al., 2014a] McNeese, M. D., Mancuso, V. F., McNeese, N. J., Endsley, T., and Forster, P. (2014a). An integrative simulation to study team cognition in emergency crisis management. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 58, pages 285–289. SAGE Publications Sage CA: Los Angeles, CA.

[McNeese et al., 2019] McNeese, N., Demir, M., Chiou, E., Cooke, N., and Yanikian, G. (2019). Understanding the role of trust in human-autonomy teaming. In *Proceedings of the 52nd Hawaii International Conference on System Sciences*.

[McNeese and Cooke, 2016] McNeese, N. J. and Cooke, N. J. (2016). Team cognition as a mechanism for developing collaborative and proactive decision support in remotely piloted aircraft systems. In *International Conference on Augmented Cognition*, pages 198–209. Springer.

[McNeese et al., 2018] McNeese, N. J., Demir, M., Cooke, N. J., and Myers, C. (2018). Teaming with a synthetic teammate: Insights into human-autonomy teaming. *Human factors*, 60(2):262–273.

[McNeese et al., 2014b] McNeese, N. J., Reddy, M. C., and Friedenberg, E. M. (2014b). Towards a team mental model of collaborative information seeking during team decision-making. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 58, pages 335–339. SAGE Publications Sage CA: Los Angeles, CA.

[Mellers et al., 2015] Mellers, B., Stone, E., Murray, T., Minster, A., Rohrbaugh, N., Bishop, M., Chen, E., Baker, J., Hou, Y., Horowitz, M., et al. (2015). Identifying and cultivating superforecasters as a method of improving probabilistic predictions. *Perspectives on Psychological Science*, 10(3):267–281.

[Mnih et al., 2016] Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., Silver, D., and Kavukcuoglu, K. (2016). Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*, pages 1928–1937.

[Mnih et al., 2013] Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., and Riedmiller, M. (2013). Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*.

[Mohammed et al., 2010] Mohammed, S., Ferzandi, L., and Hamilton, K. (2010). Metaphor no more: A 15-year review of the team mental model construct. *Journal of management*, 36(4):876–910.

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

[Neugebauer et al., 2008] Neugebauer, T., Poulsen, A., and Schram, A. (2008). Fairness and reciprocity in the hawk–dove game. *Journal of Economic Behavior & Organization*, 66(2):243–250.

[Orlitzky and Hirokawa, 2001] Orlitzky, M. and Hirokawa, R. Y. (2001). To err is human, to correct for it divine: A meta-analysis of research testing the functional theory of group decision-making effectiveness. *Small Group Research*, 32(3):313–341.

[Paolacci et al., 2010] Paolacci, G., Chandler, J., and Ipeirotis, P. G. (2010). Running experiments on amazon mechanical turk. *Judgment and Decision making*, 5(5):411–419.

[Pennachin and Goertzel, 2007] Pennachin, C. and Goertzel, B. (2007). Contemporary approaches to artificial general intelligence. In *Artificial general intelligence*, pages 1–30. Springer.

[Rosenberg et al., 2018a] Rosenberg, L., Lungren, M., Halabi, S., Willcox, G., Baltaxe, D., and Lyons, M. (2018a). Artificial swarm intelligence employed to amplify diagnostic accuracy in radiology. In *2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, pages 1186–1191. IEEE.

[Rosenberg et al., 2018b] Rosenberg, L., Willcox, G., Askay, D., Metcalf, L., and Harris, E. (2018b). Amplifying the social intelligence of teams through human swarming. In *2018 First International Conference on Artificial Intelligence for Industries (AI4I)*, pages 23–26. IEEE.

[Ross et al., 2010] Ross, J., Irani, L., Silberman, M., Zaldivar, A., and Tomlinson, B. (2010). Who are the crowdworkers?: shifting demographics in mechanical turk. In *CHI'10 extended abstracts on Human factors in computing systems*, pages 2863–2872. ACM.

[Rouse and Morris, 1986] Rouse, W. B. and Morris, N. M. (1986). On looking into the black box: Prospects and limits in the search for mental models. *Psychological bulletin*, 100(3):349.

[Salas et al., 1992] Salas, E., Dickinson, T. L., Converse, S. A., and Tannenbaum, S. I. (1992). Toward an understanding of team performance and training.

[Satopää et al., 2014] Satopää, V. A., Baron, J., Foster, D. P., Mellers, B. A., Tetlock, P. E., and Ungar, L. H. (2014). Combining multiple probability predictions using a simple logit model. *International Journal of Forecasting*, 30(2):344–356.

[Schaarschmidt et al., 2018] Schaarschmidt, M., Kuhnle, A., Ellis, B., Fricke, K., Gessert, F., and Yoneki, E. (2018). Lift: Reinforcement learning in computer systems by learning from demonstrations. *arXiv preprint arXiv:1808.07903*.

[Schaarschmidt et al., 2017] Schaarschmidt, M., Kuhnle, A., and Fricke, K. (2017). Tensorforce: A tensorflow library for applied reinforcement learning. *Web page*.

[Schraudolph et al., 1994] Schraudolph, N. N., Dayan, P., and Sejnowski, T. J. (1994). Temporal difference learning of position evaluation in the game of go. In *Advances in Neural Information Processing Systems*, pages 817–824.

[Schulman et al., 2015] Schulman, J., Levine, S., Abbeel, P., Jordan, M., and Moritz, P. (2015). Trust region policy optimization. In *International conference on machine learning*, pages 1889–1897.

[Schulman et al., 2017] Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. (2017). Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

[Sheridan and Telerobotics, 1992] Sheridan, T. B. and Telerobotics, A. (1992). Human supervisory control.

[Shirado and Christakis, 2017] Shirado, H. and Christakis, N. A. (2017). Locally noisy autonomous agents improve global human coordination in network experiments. *Nature*, 545(7654):370.

[Silver and Hassabis, 2016] Silver, D. and Hassabis, D. (2016). Alphago: Mastering the ancient game of go with machine learning. *Research Blog*, 9.

[Silver et al., 2017] Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., et al. (2017). Mastering the game of go without human knowledge. *Nature*, 550(7676):354.

[Simon and Eswaran, 1997] Simon, B. P. and Eswaran, C. (1997). An ecg classifier designed using modified decision based neural networks. *Computers and Biomedical Research*, 30(4):257–272.

[Sun et al., 2006] Sun, R. et al. (2006). *Cognition and multi-agent interaction: From cognitive modeling to social simulation.* Cambridge University Press.

[Surowiecki, 2005] Surowiecki, J. (2005). *The wisdom of crowds.* Anchor.

[Tampuu et al., 2017] Tampuu, A., Matiisen, T., Kodelja, D., Kuzovkin, I., Korjus, K., Aru, J., Aru, J., and Vicente, R. (2017). Multiagent cooperation and competition with deep reinforcement learning. *PloS one*, 12(4):e0172395.

[Tetlock and Gardner, 2016] Tetlock, P. E. and Gardner, D. (2016). *Superforecasting: The art and science of prediction.* Random House.

[Thompson, 2010] Thompson, C. (2010). Clive thompson on the cyborg advantage. *Wired Mag.*

[Tuyls and Weiss, 2012] Tuyls, K. and Weiss, G. (2012). Multiagent learning: Basics, challenges, and prospects. *Ai Magazine*, 33(3):41–41.

[Vagia et al., 2016] Vagia, M., Transeth, A. A., and Fjerdingen, S. A. (2016). A literature review on the levels of automation during the years. what are the different taxonomies that have been proposed? *Applied ergonomics*, 53:190–202.

[Van Hasselt et al., 2016] Van Hasselt, H., Guez, A., and Silver, D. (2016). Deep reinforcement learning with double q-learning. In *Thirtieth AAAI conference on artificial intelligence.*

[Von Neumann et al., 2007] Von Neumann, J., Morgenstern, O., and Kuhn, H. W. (2007). *Theory of games and economic behavior (commemorative edition)*. Princeton university press.

[Wang et al., 2016] Wang, D., Khosla, A., Gargeya, R., Irshad, H., and Beck, A. H. (2016). Deep learning for identifying metastatic breast cancer. *arXiv preprint arXiv:1606.05718*.

[Watkins, 2007] Watkins, J. H. (2007). Prediction markets as an aggregation mechanism for collective intelligence.

[Webber et al., 2000] Webber, S. S., Chen, G., Payne, S. C., Marsh, S. M., and Zaccaro, S. J. (2000). Enhancing team mental model measurement with performance appraisal practices. *Organizational Research Methods*, 3(4):307–322.

[Wellens and Ergener, 1988] Wellens, A. R. and Ergener, D. (1988). The cities game: A computer-based situation assessment task for studying distributed decision making. *Simulation & Games*, 19(3):304–327.

[Williams, 1992] Williams, R. J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256.

[Wolfers, 2009] Wolfers, J. (2009). Prediction markets: The collective knowledge of market participants. In *CFA Institute Conf. Proc. Quart*, volume 26, pages 37–44.