THE EFFECTS OF PERCEIVED COMPETENCE, PREDICTABILITY AND CONTEXT OF INTERACTION ON PERCEIVED HUMAN-LIKENESS AND HUMAN-AI INTERACTIONS

by

Stephanie Tulk Jesso

A Dissertation
Submitted to the
Graduate Faculty
of
George Mason University
in Partial Fulfillment of
The Requirements for the Degree
of
Doctor of Philosophy
Psychology

Committee:

_____ Director

_____ Co-Director

_____

_____ Department Chairperson

_____ Program Director

_____ Dean, College of Humanities and Social Sciences

Date: _____ Fall Semester 2020
George Mason University
Fairfax, VA

The Effects of Perceived Competence, Predictability, and Context of Interaction on Perceived Human-Likeness and Human-AI Interactions

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at George Mason University

by

Stephanie Tulk Jesso
Master of Arts
George Mason University, 2016
Bachelor of Science
Michigan Technological University, 2012

Director: Eva Wiese, Associate Professor
George Mason University

Fall Semester 2020
George Mason University
Fairfax, VA

DEDICATION

This is dedicated to my husband Matthew, our unborn son, and our wonderful pets.

ACKNOWLEDGEMENTS

I would like to thank the many people who have been there for me throughout this journey. My brilliant husband, Matthew, was always available to talk through the million exciting or perplexing things that came up through this research. My research assistants volunteered many hours to collect this data. In particular, Drs. Kennedy, Wiese, and Thompson were of invaluable help.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS AND SYMBOLS

Artificial Intelligence ..... AI
Human-AI Interaction ..... HAI
Human-Robot Interaction ..... HRI
Theory of Mind ..... ToM
Game Theory ..... GT
Prisoner's Dilemma Game ..... PDG
Human Beginner ..... HB
Human Expert ..... HE
AI Beginner ..... AB
AI Expert ..... AE
Reaction Time ..... RT
Analysis of Variance ..... ANOVA
Independent Variable IV
Dependent Variable ..... DV
Tukey's Honestly Significant Difference ..... Tukey's HSD
Multivariate Analysis of Variance ..... MANOVA
Chi Squared ..... χ^2
Standard Deviation ..... SD
Partial Eta Squared ..... ηp^2
Reaction Time ..... RT
Don't Starve Together ..... DST
Mechanical Turk ..... mTurk

ABSTRACT

THE EFFECTS OF PERCEIVED COMPETENCE, PREDICTABILITY, AND CONTEXT OF INTERACTION ON PERCEIVED HUMAN-LIKENESS AND HUMAN-AI INTERACTIONS

Stephanie Tulk Jesso, Ph.D.

George Mason University, 2020

Dissertation Director: Dr. Eva Wiese

As robots and AI are designed for social interactions and incorporated into the world, it is important to examine the extent to which humans can perceive such artificial agents as humanlike based solely on the observation of their behaviors. The while psychologists, neuroscientists, computer scientists and designers have studied related phenomena for many years, more work is required to develop a deep understanding of how people perceive the actions of AI compared to humans within a complex, interactive environments, and how this varies as a function of the AI's competence in navigating that environment.

The experiments outlined in this dissertation attempt to clarify how individuals make distinctions of human-likeness on the basis of observable behavior, how the perception of human-likeness is related to the perceptions of competence at a given task as well as how predictable or explainable these behaviors are perceived to be, and finally

how the context in which these judgements are made contributes to expectations and overall perceptions of a complex mind.

The results from these experiments suggest (1) that increased competence at a task is accompanied with an increase in perception of human-likeness, (2) the relationship between perceived predictability/explainability and human-likeness depends on how competent an entity is perceived to be and (3) the context of an interaction can lead to different expectations of human-likeness and can potentially affect the overall relationship individuals develop with human and AI entities.

INTRODUCTION

Differentiating Human and AI Actions

While the fields of Human-AI interactions (HAI) and Human-Robot Interactions (HRI) are still developing today, human perception of non-human agents has been a topic of interest to multiple scientific fields for many years. While the research has yielded many insights, some major questions remain that are pivotal to our understanding of how to design effective HAI and HRI.

Anthropomorphism and Mind Perception

Some have found that only devices with very "humanlike" physical appearance can elicit humanlike social interactions from human interaction partners (MacDorman & Ishiguro, 2006). If an agent is perceived as having a mind, as humans are by default (Epley, Waytz & Cacioppo 2007; Gray, Gray & Wegner, 2007), individuals may adopt the "intentional stance" (Dennett, 1989), or the conscious belief that the agent has a "mind of its own" that is capable of rational, intentional thoughts and executing actions on its own accord. This top-down belief can affect how attentional resources are deployed so that the person is sensitive to subtle social cues such as gaze direction (Wykowska et al., 2014; Caruana et. al., 2017) which, in the context of many social interactions, can help people communicate important information effectively (Frischen, Bayliss & Tipper, 2007; Mutlu et al., 2009).

The perception that a non-human agent has a "mind of its own" can be triggered by perceived similarity to humans through observations of humanlike appearance and motion, personal motivations like a need to explain negative or evil acts, but also the observation of behaviors that appear to be independent (see Waytz et al., 2010a for a review). The types of behaviors artificial agents exhibit, e.g., physical characteristics of a robot's motion (Wykowska et al., 2015; Bisio et al., 2014), executing goal directed actions (Gazzola et al., 2007), engaging in humanlike activity such as eye contact (Kompatsiari et al., 2019), joint attention (Pfeiffer et al., 2011), or acting unpredictably as if of their own volition (Short et al., 2010; Hayes et al., 2014; Waytz et al., 2010b; Salem et al., 2013), and the context an agent's actions are presented in (e.g., competitive or cooperative: Pfeiffer, 2011) can have strong impacts on the overall perceptions of these agents.

Competence and Anthropomorphism

There seems to be a gap in the literature related to how skill or competence at a task affects the perception of human-likeness. Humans seem to assess competence in others regularly and universally, and the perception of competent elicits positive attribution and social treatment when the entity is also perceived as good-natured (Fisk, Cuddy & Glick, 2006). The perception of competence may be related to the perception of "agency", or the capacity to plan and act of one's own volition (Waytz et al., 2010a). While Waytz, Heafner & Epley (2014) found that the attribution of human-likeness could increase the trust that an agent was competent, they did not directly manipulate competence to measure the effect on human-likeness.

While competence may be an expectation of human-like entities, perfection may

be a default expectation of AI. The "perfect automation schema" has been theorized after examining peoples' use of automated decision aids (Dzindolet et al., 2002). Dzindolet and colleagues found evidence that participants had a pre-existing assumption that automation should provide near-perfect information, and that those who were warned that the automated aid was not perfect used it more than those who were not warned. Merritt et al. (2015) expanded upon the theory and demonstrated that the perfect automation schema may be more related to a perception of all-or-none (i.e., automation works perfectly or not at all) rather than an assumption of very high-level, competent performance.

A gap in our current understanding is how people will evaluate advanced AI that is capable of human-level performance on complex tasks. Some examples have emerged in the past few years and have excelled at tasks that humans once believed that only a human could accomplish (Mnih et al., 2015; Silver et al., 2017; Vinyals et al., 2019; McKinney et al., 2020). When it comes to the evaluation of AI that can perform competently at complex social tasks, observations of their movements and actions will likely affect overall perceptions of mind.

Perception of Animate vs. Inanimate Objects

A human's ability to distinguish between movements produced by animate and inanimate objects is developed in infancy (Rakison & Poulin-Dubois, 2001). Rakison and Poulin-Debois theorized that infants pay attention to seven different characteristics of motion to help them make such determinations, including different qualities of the movement itself, if motion has a purpose (goal directed or not), and if the motion was intentional or

accidental. Biological motion, or motion that is produced by biological organisms from their organic musculoskeletal systems, may also provide a strong indication of animacy that helps infants learn distinctions (Poulin-Dubois, Crivello & Wright, 2015). An infant's sensitivity to biological motion may also be an early stage in their development of Theory of Mind (Frith & Frith, 1999).

Sensitivity to biological motion is a social mechanism that helps humans sense actions and intentions from observing movement (Blakemore & Decety, 2001). The extent to which activation in social brain areas differs when observing human and robots/AI actions has been studied extensively (for a review, see Wiese, Metta & Wykowska, 2017). Some research has demonstrated that the human action-perception system is similarly sensitive to actions performed by humans and mechanistic robots (Gazzola et al., 2007, Bisio et al., 2014). However, others have shown that non-human social agents do not activate the more complex social brain areas to the same extent that human interaction partners do (Takahashi et al., 2014; Wang & Quadflieg, 2015; Sanfey et al., 2003). However, the technological advancement of the artificial agents used in these studies is still far from human-level competence, and more work is needed to adequately trigger the perception that such agents have complex mental states (Wiese, Metta & Wykowska, 2017). Understanding human expectations for robot behavior can help guide the development of cognitive models that can be implemented in such agents to better align them for desirable HAI and HRI in the real world (Breazeal & Scassellati, 1999; MacDorman, 2006). It is still necessary to conduct further research to understand how humans will perceive the actions that AI agents carry out to accomplish their tasks, and

whether or not it induces the belief that such actions were carried out thoughtfully by an entity that possesses a mind that thinks and makes decisions like humans do, or simply as machines carrying out mechanical actions.

Theory of Mind

While infant humans can learn to classify and understand objects around them based on motion, our broader understanding of other people is complex and our need for understanding other people goes beyond discerning simple motion and actions. Theory of Mind (ToM) entails the attribution of complex mental states onto other entities and allows humans to perceive others as having their own thoughts, emotions, beliefs, desires and intentions as a mechanism to understand and predict their behaviors (Premack & Woodruff, 1978; Baron-Cohen, Leslie & Frith, 1985). An early psychological experiment related to ToM was Heider and Simmel's 1944 Study of Apparent Behavior. It showed that humans ascribed complex mental states and intentions to moving simple shapes after watching animated vignettes where the shapes moved around the screen (Heider & Simmel, 1944). This stimuli was later adapted to study ToM with neuroscientific methodology (Castelli et al., 1999; Martin & Weisberg, 2003). When the vignettes of moving shapes were interpreted as depictions of social interactions, participants had greater neural activity in social and emotional brain areas, compared to when vignettes were interpreted as mechanical actions, where greater activity was observed in regions associated with identifying usable tools (Martin & Weisberg, 2003). Importantly, this interpretation did not rely on human morphology or biological motion, indicating that even without other

outwardly human-like qualities, robots or AI that appear to act in social ways may be able to trigger ToM in humans.

It is no surprise that roboticists have focused on the development of social robots and AI that can trigger ToM in humans (Brezeal and Scassellati, 1999) and have their own ability to represent others with ToM (Scassellati, 2000; Rabinowitz et al., 2018). The purpose of developing robots that can exhibit and trigger ToM is to allow for easier interpretation of social actions and better human-robot/AI relationships (Brezeal and Scassellati, 1999; Scassellati, 2000). If robots and AI are developed with sufficient ToM-triggering abilities, it is possible that they will be seen as entities with intelligence, and it may become difficult to distinguish between such entities and humans in certain tasks, yet much more work remains to understand how to develop robots and AI that can reliably trigger ToM.

The Turing Test

One method for investigating how distinct or similar AI-produced social actions are to human-produced actions is to evaluate how accurately individuals can distinguish between these. Such an approach provides both an indication of how similar an agent's observable behavior is to our expectations of human behavior and an indication of how "intelligent" or competent an artificial agent is perceived to be at a certain task. Alan Turing, father of modern computers, described a test in which a human would evaluate a machine in five minutes of unrestricted conversation and decide whether they believed they were communicating with an actual human or a machine. If 30% of judges were convinced the machine was a human, it would have passed the Turing Test and should be considered

"intelligent", as they could display human-level competence in a conversation (Turing, 1950). While this typical Turing Test tests the conversational capacity of AI, a few studies have conducted the test on the basis of behavior (Pfeiffer et al., 2011; Wykowska et al., 2015; Osawa et al., 2012; Tulk et al., 2018). Results indicate that humans are sensitive to true human-likeness in low level behaviors and can distinguish between human- and AI-controlled arm motion with above chance accuracy (Wykowska et al., 2015), but the context of the interaction affects the way the behavior is judged such that people have different expectations of humanlike behaviors in cooperative, competitive, and naive contexts (Pfeiffer et al., 2011). Judging accurately is still a difficult task (Osawa et al., 2012; Tulk et al., 2018), and people may have pre-existing assumptions about what robotic movements look like (Wykowska et al., 2015). Importantly, these behavioral Turing Tests involved AI with simplistic cognitive abilities and performance, so they cannot provide a conclusive understanding of how a complex agent that has internal thoughts, beliefs and intentions will be evaluated or perceived. There is still a need for research involving complex and Social AI to better understand how perception of human-likeness is affected by the presence of complex cognitive abilities in AI.

Human-AI Interactions

To fully understand Human-AI Interactions (HAI), it is important to know the extent to which people will form relationships with AI that resemble human-human relationships over the course of interactions, including how interested humans are in cooperating or competing with AI entities when they have the freedom to decide for themselves.

Game Theory and Social Decision Making

Deciding how to treat others across different social contexts is a complex process that involves evaluations of the interaction partner, considerations of what each party has to gain or lose, and the relationship between entities (Lee, 2008). It is unsurprising that social decision making involves social and emotional brain areas that respond differently depending on the level of mind that is perceived in an interaction partner (Takahashi et al., 2014; Sanfey et al., 2003). To explore this further, many researchers have used game theory to help elucidate the many complexities of social decision making. Game theory is a mathematical framework that allows for the examination of motivations and strategies within interactions between different actors (Ross, 2001). Historically, it has been used to understand cooperative and competitive relationships in economics (Axelrod and Hamilton, 1981; Miller, 1996), evolutionary biology (Trivers, 1971; McNamara et al, 2008) and computer science (Leibo et al., 2017), but also as an experimental tool to study social decision making in psychology (Sally, 1995) and neuroscience (Sanfey et al., 2003; Lee, 2008; Takahashi et al., 2014).

A common social phenomenon studied in game theory experiments is reciprocity in cooperative or competitive actions, where kindness is returned and unfairness is punished. Humans have evolved with a strong expectation and tendency towards reciprocity in social interactions (Gouldner, 1960). This can be demonstrated in one of the most popular game theory games called the Prisoner's Dilemma Game (PDG, as seen in Figure 1), in which the two actors must consider strategy in the presence of greed and fear.

While mutual cooperation ensures a better outcome than mutual competition, it requires mutual trust because each actor has more to gain one-way competition (i.e., you compete and your partner cooperates) which induces the greed for a better outcome as well as the fear of being taken advantage of if your partner competes while you cooperate. While a traditional economic view of the interaction predicts the outcome of mutual defection, when people actually play the PDG in sequential games, mutual cooperation is often achieved through reciprocity (Axelrod and Hamilton, 1981). Tit-for-tat, or the choice to act with perfect reciprocity in sequential games or interactions, is a social strategy that enforces mutual cooperation by showing an interaction partner that their best choice is to cooperate, as competing is punished in subsequent rounds to a degree that eliminates any winnings gained by competing (Axelrod & Hamilton, 1981). Humans have also been observed using tit-for-tat with robot interaction partners just as they would with human partners (Sandoval et al., 2015), though it is inconclusive whether or not this tendency was instinctual or meant to display social information to the robot agents.

In addition to the PDG, there are many other game theory games that give insights into other types of social interactions. The numbers depicted in a game theory matrix (see Figure 1) represent player motivations and allow for the comparison of social context in addition to understanding outcomes. For instance, the context of the PDG is very different from the Stag Hunt game, in which two hunters decide whether or not to work together to take down a stag or work separately to catch a rabbit each. Since each hunter knows that they cannot capture a stag alone, and that the meat they gain from a stag is more than the

meat from a rabbit, mutual cooperation is a natural result; see Figure 1. When compared to a representation of global nuclear deterrence, or the game of Mutual Assured Destruction, neither actor has much to gain from mutual cooperation, but is extremely motivated to avoid mutual competition for fear of catastrophic destruction and an end to human civilization; see Figure 1. While each of these three games can result in mutual cooperation, the motivations and social context around each game is very different. An additional benefit to the mathematical representation is that it provides a simple way to translate between social context and machine language. Empirical game theoretic analyses, where matrix weights and corresponding games are derived from empirical performance data, can be used to understand the social dynamics in multi-agent systems (Wellman, 2006; Leibo et al., 2017). People in the field of HRI/HAI have advocated a game-theoretic approach to understanding strategies in decision making during HRI/HAI (Lee & Hwang, 2008) and for developing approaches for training AI to have satisfying and appropriate relationships with humans (Hadfield-Menell et al., 2016; Palaniappan et al., 2017).

A gap in current literature is an understanding of how individuals choose to interact with AI without any motivational influence. Research on true HAI within games have fixed the interactions to strictly cooperative (Ehsan et al., 2018) or competitive interactions (Silver et al., 2017; Vinyals et al., 2019), which may be useful for understanding how far AI has advanced, but not how humans naturally perceive interactions or how humans might choose to develop these relationships on their own over the course of interactions.

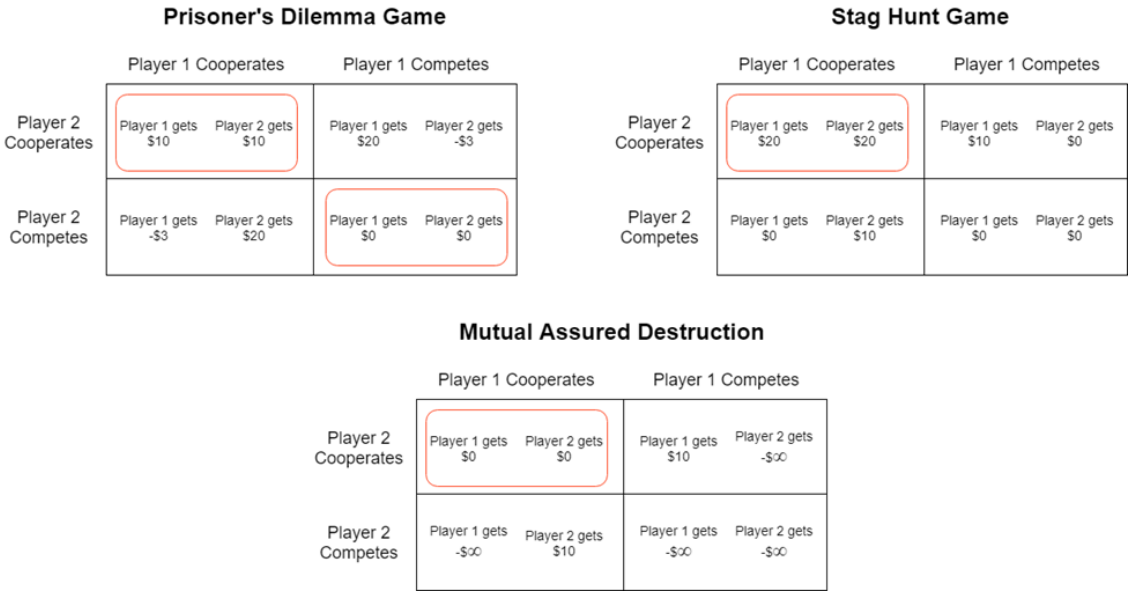

Figure 1: The Prisoner's Dilemma Game (PDG), Stag Hunt Game, and Mutual Assured Destruction. In the PDG, two players must decide whether to cooperate or compete with one another. While more can be gained from mutual cooperation, the temptation to compete and the fear of being taken advantage of by another player can lead to mutual competition and a worse outcome. While mutual competition is predicted by a traditional economic standpoint, mutual cooperation is often reached when actors play repeated PDGs with a tit-for-tat strategy. Unlike the PDG, both players in the Stag Hunt Game gain the most from mutual cooperation than any other strategic decision. In the Mutual Assured Destruction Game, while both parties have nothing to gain from mutual cooperation, they have everything to lose from mutual competition.

HAI in Videogames

When investigating how humans interact with AI, it is important to create AI that has the ability to respond to humans on its own without using "Wizard of Oz" techniques to fake a response because people's perceptions of agents are affected by subtle cues such as timing of a response or movement (Epstein, Roberts & Beber, 2009; Wykowska et al., 2015) or mode of interaction (Short et al., 2010). Videogames can provide an environment to study how humans interact with AI socially because they are already developed for rich social interactions, and many provide the opportunity to make custom modifications to game code which can allow researchers to build systems to capture behavioral data from within the game. Videogames have been used as a platform for training and evaluating AI

performance (Laird & VanLent, 2001; Mnih et al., 2015), and to investigate how humans perceive human and AI performances differently by measuring human behavior and subjective experience (Tulk et al., 2018; Ehsan et al., 2018) as well as physiological measures (Lim & Reeves, 2009).

Research Questions

While there is a strong academic history of studying human perceptions of robots and Simple AI, it is apparent that gaps remain in our understanding of how humans will perceive competent AI that can exhibit human-level performance.

It is not well understood currently if the actions of competent AI will be perceived as purely mechanical actions carried out based on human-made programming, or as a result of decisions made by intentional minds with complex inner states. If these perceptions are non-binary and occur on a spectrum, more work is needed to understand the features of that spectrum. While humans have been shown to perceive complex ToM states while knowingly watching the actions of non-human agents, there is still a lack of understanding of how the perception of ToM states can be triggered behaviorally or as a result of non-verbal interactions with such AI. Finally, more work is needed to understand how humans are likely to treat such agents socially, if they will naturally be regarded as collaborators or competitors, and what types of behaviors and qualities can affect this relationship.

Understanding these gaps in the current literature lead to the following research question:

1. How do people differentiate between actions produced by other humans and "mechanical" actions produced by AI?

After a pilot study (i.e., Experiment 1) was conducted to answer question 1, 2 additional questions came into focus:

2. To what extent do competence and predictability contribute to the perception of human-like behavior?

3. What characteristics of social interactions do humans use to differentiate between humans and AI, and how do Human-Human Interactions and Human-AI Interactions (HAI) differ within a complex environment?

The following experiments have been developed to answer these questions.

Research

To develop a fuller understanding of human perceptions of AI and how it affects HAI, three experiments have been conducted in which participants observed and/or interacted with AI of different levels of competence within popular videogames, Super Mario World and Don't Starve Together. Experiments 1 and 3 featured a behavioral Turing Test, where participants judged the identity of human and AI agents and were interviewed about what behavioral cues led to these decisions. Experiment 2 had participants respond with the extent to which they perceived agents as humanlike based on observable behaviors on 10-point Likert scales rather than a Turing Test. In all experiments, additional measures were recorded to better understand participants' overall perceptions of agents related to their actual and perceived identities.

EXPERIMENT 1: MARI/O

How do people differentiate between actions produced by other humans and "mechanical" actions produced by AI?

The first experiment was developed as an exploratory pilot study. The purpose of the first experiment is to characterize how people differentiate between actions produced by other humans and actions produced by AI. Specifically, this experiment was designed to answer question 1: *how do people differentiate between actions produced by other humans and "mechanical" actions produced by AI*. Participants watched pre-recorded vignettes of the game Super Mario World played by human and AI players. The human players included a total beginner who had never played any 2D Mario game prior, and an experienced player who had played the game many times throughout his life. The AI included an early-generation and late-generation of a genetic algorithm / neural network (Stanley & Miikkulainen, 2001; adapted for Mario by Youtuber Seth Bling, Bling, 2015). Participants watched vignettes of each agent and rated them as beginners/experts and human/AI, then responded to a structured interview about their ratings and justifications for those ratings.

Participants

Participants include 27 undergraduates recruited from George Mason University's Sona system (14 female, average age 20.6 years, SD = 1.7). On average, they spent 5.4 hours on a computer each day (SD = 2.6 hours). The only screening criterion was that participants be over the age of 18.

The primary data of interest in this study was the qualitative data generated by the interviews, therefore a sample size of 27 participants was considered sufficient. Often, anywhere from 5 - 50 participants are collected for studies that involve in-depth interviews (Dworkin, 2012). Each participant was asked 6 questions for each of the 4 agents and 2 additional questions, yielding 702 natural language responses. A post hoc power analysis was conducted using G*Power to determine the power for the ANOVAs used to analyze quantitative data given the number of participants used in this study. With a medium-large effect size $f = 0.35$, α error probability of .05, the observed power = 0.41.

Stimuli

Four different agents were selected for the Turing Test: Human Beginner (HB) and Expert (HE), and AI Beginner (AB) and Expert (AE). Each played the game Super Mario World on a PC via the BizHawk emulator. Human recordings came from two undergraduate research assistants: one who had never before played any 2D Mario game (HB), and one who claimed to have had a lot of prior experience with Super Mario World throughout his life (HE). The AI sequences were generated by recording the performance of an agent that learned how to play the game through a NEAT (NeuroEvolution for Augmenting Topologies) genetic algorithm (Stanley and Miikkulainen, 2002). This agent was based on Seth Bling's original MarI/O implementation of the NEAT algorithm (Stanley & Miikkulainen, 2001). The genetic algorithm used each genome's fitness and random mutations to "breed" new generations that eventually learned how to play the game. An adaptation was created to include total coins collected and overall game score in

the calculation of fitness in order to more closely resemble a human's motivation when learning to play the game. This addition resulted in more complex behaviors; for instance, the AI would sometimes move backwards (right to left) to jump on enemies or collect coins that were missed, whereas the original would never move backwards. Performance was recorded over the course generation of neuroevolution, with performances from an early generation constituting the Beginner AI (AB), and the final generation constituting the Expert AI (AE). Video vignettes were made for the 4 different agents, each playing through the first level of Super Mario World. Each agent's performance was recorded and subdivided into individual vignettes, each starting at the beginning of the level and terminating either when Mario died or finished the level. All recordings were first screened and vignettes were selected to exhibit the full range of each agent's behaviors and errors. No sound was recorded. The total number of vignettes for each agent were: 11 for AE, 11 for AB, 8 for HB, 10 HE, resulting in approximately 5 total minutes of video length for each agent.

Apparatus

Participants watched videos on a 19-inch ASUS VB Series VB198T-P monitor with a 4:3 aspect ratio and refresh rate of 60 Hz, sitting approximately two feet away from the monitor, and indicated their responses through a keyboard. Google Forms was used to administer a basic demographic survey and record responses to interview questions. Stimuli videos were recorded using OBS screen recording software. Adobe Premiere was

later used to cut videos into vignettes and crop the display such that only the game screen was visible for all vignettes.

Measures

Participants filled out a basic demographic questionnaire including their gender, age, education level, average time spent on a computer per day and average time spent playing videogames per week.

After each trial, participants' decisions about an agent's identity (humanness: human or AI, and expertise: beginner or expert) and Reaction Times were recorded through keyboard responses. The eight interview questions are presented as follows:

1. What is your experience and approximate skill with Mario?
2. What agent do you think you were watching? Why? (Asked after blocks 1-4)
3. What features of the performance made you think it was (Expert/beginner)? (Asked after blocks 1-4)
4. What features of the performance made you think it was (Human/AI)? (Asked after blocks 1-4)
5. What do you think the agent's goal was while they were playing? (Asked after blocks 1-4)
6. What made you think it wasn't (Expert/Beginner)? (Asked after blocks 1-4)
7. What made you think it wasn't (Human/AI)? (Asked after blocks 1-4)

8. If you were going to play a game with one of the agents, which one would you prefer? Would it be cooperative or competitive, and why?

While the interview questions were preplanned, research assistants would often ask follow-up questions to encourage participants to elaborate on their responses to collect as much qualitative data as possible. Interview responses given after each block were transcribed by a research assistant to record the final explicit decision of each agent's identity (the Turing Test/humanness and expertise) and the verbatim natural language participants used to describe agents while answering each question. After completing all blocks, participants were asked what strategy (cooperative or competitive) they would use if they were to play a game with one of the 4 agents in the future and justifications for this decision.

Procedure

Participants were instructed that they would be performing a series of Turing Tests after observing the 4 agents (HB,HE,AB,AE) play Super Mario World, in which they would need to indicate if they believed the agent was a human or an AI, and if the agent was a beginner or expert. Super Mario World is a 2 dimensional platformer videogame in which players collect points and avoid enemies as they move through the level. Game actions are limited to moving right, left, jumping, ducking, and spinning while jumping. Players must perform these actions at appropriate times to gain points and avoid being killed by enemies or other obstacles, all the while moving towards the right of the screen

to complete the level. The player has a limited amount of lives to expend before a game over, which can be seen in the upper left corner of the screen. The player can also see how much time they have remaining to complete the level and the amount of coins and points they have earned in the upper right on the screen. An example view of the game can be seen in Figure 2.

Figure 2: A Typical View of the Game Super Mario World.

Participants were told that they would watch a series of performances (vignettes) from one of the four agents, presented at random. Within each block, participants watched all vignettes rated the player as a beginner or expert, and human or AI after each on trial by key press on a keyboard. To indicate their response, participants positioned their hands on a keyboard with their index fingers on the "w" and "o" keys, and their thumbs on the "x"

and “m” keys. On the left hand, a key response of “w” corresponded to the belief that the agent’s identity was HE, and response of “x” corresponded to the belief that the agent’s identity was HB. On the right hand, “o” corresponded to AE, and “m” to AB. As Reaction Time was recorded for analysis, participants were instructed to keep their hands in position on the keyboard for the entire block so they were ready to respond when prompted. Once a participant finished a block, they were verbally interviewed about their overall perceptions of the agent and their final explicit belief of whether the agent was a human.

The trial sequence within each block was as follows: participants watched a vignette of the agent playing the first level of Super Mario World, selected at random from the full list of the agent’s performances. Prior to the presentation of the vignette, a fixation cross was presented for 500ms. After the vignette ended, a screen was presented asking participants to give their response through the keyboard. Once participants responded, a new trial would begin, and this cycle continued until participants had watched and responded to each vignette within the block.

After participants finished each block, the researcher conducted a brief interview to understand participants’ final decision on the agent’s identity (humanness and expertise), what performance features and overall qualities led them to this decision, and depending on their decision on humanness, what made them think it was not the other agent type (i.e., interview questions 2-7). Finally, after completing all four blocks, participants were asked about their cooperative and competitive preferences if they were to play a game as a co-player with any of the agents and what factors influenced that decision. The experiment took about 30 minutes to complete.

Quantitative Analysis

Performance on the Turing Test was evaluated by comparing the relative frequency that players were rated as humans. T-tests were used to compare accuracy in detecting humanness against chance on individual trials (similar to Wykowska et al., 2015) to estimate how sensitive participants were to humanness in the experiment.

Participants' keyboard responses indicating beliefs of humanness and expertise after each trial were subjected to two 2 (Humanness: human vs. AI) x 2 (Expertise: beginner vs. expert) ANOVAs. A Linear Mixed Effect (LME) model was used to analyze reaction times for each trial with respect to the agent's Humanness and Expertise.

Qualitative Analysis

Natural language responses to interview questions were recorded through transcription and coded and analyzed by two raters to determine how people described their perceptions of each agent, including what qualities are commonly associated with humans and AI (when considering true and perceived identity), what qualities led to accurate and inaccurate identification of agents, and how language varied with expertise.

The qualitative analysis was a multi-step process. First, categories were developed for coding the interviews. This was first done in a bottom-up fashion using Grounded Theory (Suter, 2012, Glaser, 1998; Glaser & Strauss, 1967), where interview data was first reviewed without pre-existing expectations of categories. Interview questions 2, 4, and 7 were focused on during the qualitative analysis as these were most related to the research question. While reviewing interview data, attention was paid to themes that seemed to

emerge in order to determine if these occurred with enough frequency to include them in the overall categories. After first reading through all the data, coming up with categories ad hoc, and conversations with my doctoral advisor, a nearly final set of categories was decided upon, and all interview data from the relevant questions was then coded.

To calculate inter-rater reliability, a consensus approach was adopted between two raters, in which raters first worked independently to rate all interview data from the relevant questions, then reconciled differences to reach a greater consensus by reconsidering all instances where coded texts were inconsistent between the two (Syed & Nelson, 2015). Two raters used the near final set of categories and began coding interviews. After an initial pass, minor modifications to the categories were adopted (splitting or combining concepts) to arrive at the final set of categories. RQDA (R Qualitative Data Analysis; Huang, 2016) was then used to code interviews. In RQDA, the raters read through interview data text and could highlight specific text with the cursor and mark it with any category. In this way, the actual phrases each rater associated with each category were saved, as well as the overall frequency of categories, and the whole processed interviews that indicated how and which sections were coded. After the raters finished coding interviews, a python script was written to compare the consistency of codes by listing out all statements that were coded by one or both raters. In order to adopt a consensus, each rater was given a copy of the python output to determine if they agreed or disagreed with the other rater when there was inconsistency, as well as to re-examine their own inconsistent codes. Cohen's Kappa was calculated as a measure of Inter-rater reliability based on the final decisions, and a $\chi 2$ between raters was calculated to show the likelihood of agreement between the raters.

Finally, an empirical game theoretic analysis was performed on participants' responses to their strategic preferences for future interactions with the agents as a representation of the perceived social context between actors (similar to Wellman, 2006 and Leibo et al., 2017). The empirical game theoretic analysis was based on an analysis of stated strategic preferences and justification of those preferences in order to determine what game theory game best represented the relationship. First, relative approximations of game theory matrix weights were determined from participants' statements. Figure 3 shows a matrix with variables representing the payoffs for each player given the outcome. The participants' preference was given the highest value, and other relative values were determined based on accompanying statements. Symmetry (i.e., players 1 and 2 have the same payoffs for self/other decision pairs, A = B, C = F, D = E, G = H) is generally assumed.

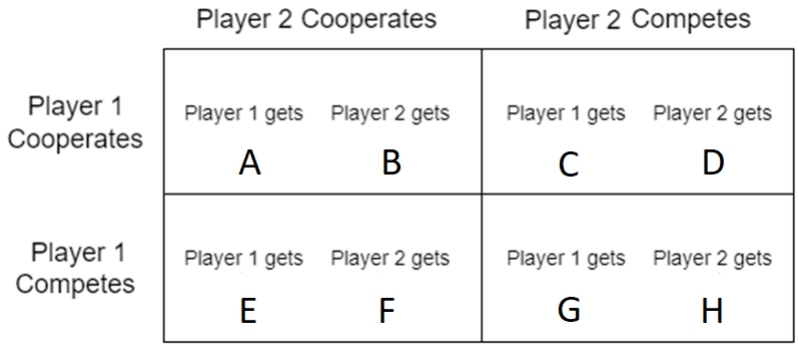

Figure 3: Game Theory Matrix with Variables Representing Individual Payoffs. The first position in each cell represents player 1's payoff and the second position represents player 2's payoff.

For example, if a participant said they preferred cooperation, and the justification is that they assumed both players would perform better if they worked towards the same goal, it implies symmetry because they prefer that player 2 also cooperates, and each has more to gain from mutual cooperation than all other options. That means that A = B > C,D,E,F,G,H. It also implies that utility

in the game (e.g., survival, score) is important to the player, so options in which player 2 competes while player 1 cooperates should be worst, as this situation would likely yield the worst score, so C < A,B,E,G,H. Since the participant did not mention that they were fearful that they could be taken advantage of or that either has something to gain from tricking their partner into cooperation while choosing to compete, it doesn't suggest that there is a strong difference between D and H, and since symmetry is assumed, D = F = G = H, so A = B > D,F,G,H > C = E. This configuration is consistent with a Stag Hunt. Importantly, a Stag Hunt is a coordination game with two pure strategy Nash Equilibria, meaning that each player is better off knowing what the other is doing when making their decision, and better outcomes are achieved by choosing to coordinate (Ross, 2001).

Similarly, if a participant expresses that they prefer to compete for the challenge and for enjoyment, it implies that they prefer if player 2 competes too, as this leads to maximum challenge and enjoyment, so symmetry is assumed. Since their preference is for mutual competition, achieving a higher score than player 2 is important, so C is again the worst outcome. Again, the lack of strong evidence that there is fear of being taken advantage of, so D is approximately equal to H, and again D = F = G = H, so G = H > A,B,D,F > C = E. This type of configuration is consistent with a coordination game, and is similar to a Stag Hunt, except with a reversed preference of mutual competition over cooperation (Ross, 2001).

Conversely, a participant saying they would compete because they assume they would easily overtake player 2 implies that the outcome for choosing competition is always better than cooperation, so E,H > A,C. If beating player 2 is the desired outcome, this would be even easier if player 2 cooperates, so E > G. Since winning is preferred, the worst outcome is C, therefore E>G>A>C, which is consistent with a Deadlock. Unlike a coordination game, Deadlock has 1 pure strategy Nash Equilibrium, where a player has no

incentive to coordinate with the other player and does best when choosing competition (Ross, 2001).

Results

Quantitative Results

On the Turing Test, human players were explicitly rated as humans greater than 70% of the time (84%, SD = 36.7%, for human beginner, 74%, SD = 43.8%, for human expert), suggesting that their performances were perceived as humanlike. Similarly, the AI beginner was rated an AI 78% of the time, SD = 41.6%, in explicit responses. However, the expert AI was explicitly rated as a human 58% of the time, SD = 49.4%, indicating that this agent was perceived as a human more often than correctly identified as an AI. Importantly, some participants were uncertain of an agent's identity and did not make an explicit report. In terms of rated expertise, all agents we labeled as beginners or experts received the same explicit rating by greater than 70% of participants, validating that their expertise was in general perceived as intended.

For keyboard responses on individual trials, t-tests revealed that accuracy in detecting Humanness was significantly above chance for the human beginner (63% accurate, $t(215) = 4.09$, $p < .001$), human expert (64% accurate, $t(268) = 4.89$, $p < .001$), and AI beginner (62% accurate, $t(296) = 4.49$, $p < .001$), whereas accuracy in detecting Humanness was nearly significantly below chance for AI expert (45% accurate, $t(295) = -1.63$, $p = .052$), all one-tailed.

Average percent of trials rated as human were significantly impacted by Humanness ($F(1, 104) = 10.00$, $p < .001$, $\eta p2 = .088$) but not Expertise ($F(1, 104) = 2.66$, $p = .106$, $\eta p^2 = .025$); the Humanness x Expertise interaction ($F(1, 104) = 2.27$, $p = .135$, $\eta p^2 = 0.021$) was also not significant.

As a post hoc analysis, a one-way ANOVA was conducted with player as an Independent Variable (IV) and average ratings of humanness in trials as a Dependent Variable (DV), which was significant ($F(3,104) = 4.98$, $p = 0.003$). A Tukey's Honestly Significant Difference (HSD) showed that the differences between the HB (mean = 63.4%, SD = 24.1%) and AB (mean = 37.3%, SD = 36.6%) was significant ($p = 0.007$), as was the difference between the HE (mean = 64.3%, SD = 23.7%) and AB ($p = 0.005$). No other differences were significant. Results can be seen in Figure 4.

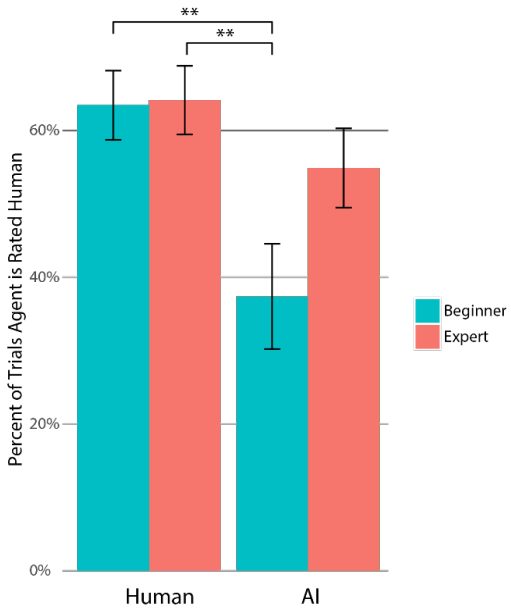

Figure 4. Percent of All Trials Each Agent is Rated as Human (Response of HB or HE) Across All Participants. Both human agents and the expert AI agent were rated as human significantly more often than the beginner AI.

Humanness significantly impacted the average percent of trials in which players were rated as experts ($F(1, 104) = 31.99$, $p < .001$, $\eta p^2 = 0.235$), as did Expertise ($F(1, 104) = 90.31$, $p < .001$, $\eta p^2 = 0.465$). The interaction was also significant ($F(1, 104) = 4.12$, $p < .05$, $\eta p^2 = 0.038$). A post hoc Tukey's HSD showed that there were significant differences between the HB (mean = 36.6%, SD = 28.3%) and the AB (mean = 0.7%, SD = 2.4%; $p < 0.001$), the HE (mean = 71.4%, SD = 29.6%) and the AB ($p < 0.001$), the HB and AE (mean = 54.4%, SD = 24.3%; $p = 0.038$), the HE and HB ($p < 0.001$), and the AB and AE ($p < 0.001$). The difference between the HE and AE was nearly significant ($p = 0.056$). Results can be seen in Figure 5.

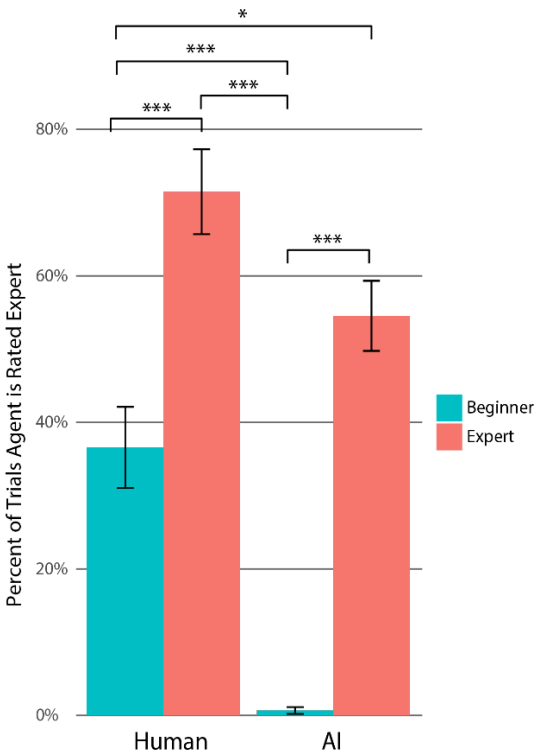

Figure 5. Percent of Trials that Each Agent was Rated as an Expert. All agents' ratings are significantly different from one another.

Average Reaction Times (RTs) can be seen in Figure 6. A LME model was created to control for repeated measures and investigate how Humanness and Expertise affected reaction times throughout trials. An ANOVA was then conducted on the model to compare the effects of Humanness and expertise. Humanness significantly affected RT ($F(1, 1048) = 26.92$, $p < .001$), as did Expertise ($F(1, 1048) = 11.27$, $p < .001$). The interaction was also significant ($F(1, 1048) = 25.34$, $p < .001$).

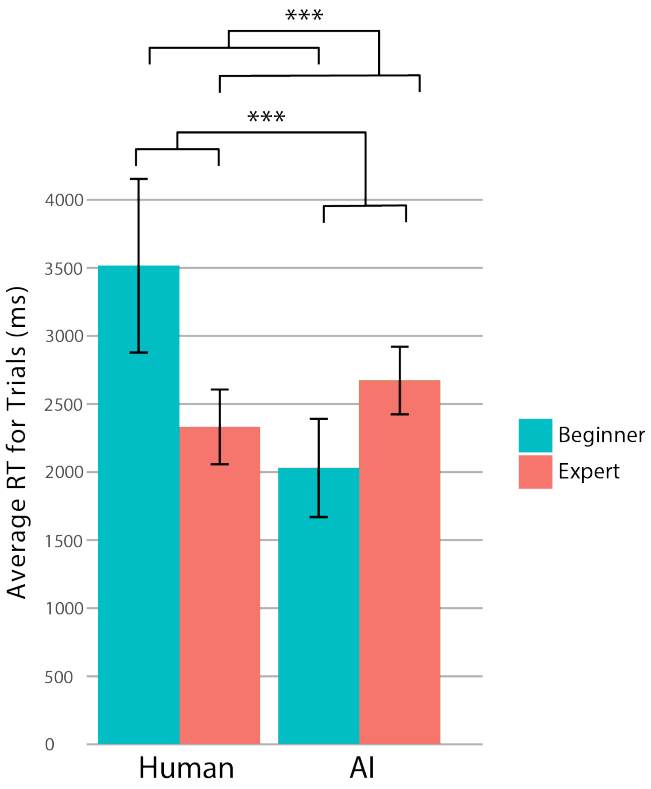

Figure 6: Average Reaction Times (In MS) of Responses for Each Agent. Humanness and Expertise significantly affected reaction times. The interaction was also significant.

Qualitative Results

From the qualitative analysis, 324 responses (3 interview questions * 4 agents * 27 participants) were analyzed resulting in 19 unique categories and 550 coded statements. Three additional categories (used Yoshi, had personality, and was pre-programmed) were coded, but the coded data were either unreliable or did not add any value to the analysis, and were removed. Each category, count, percent of participants who mentioned the code at least once, Cohen's Kappa, and χ^2 is presented in Table 1. After the initial categorization, mid-level and overarching categories were created by examining how the categories related to each other in the context of the participants' task (i.e., axial coding; Suter, 2012). The mid-level categories were related to how much top-down information was used when the observations were made. Some mid-level categories were associated with bottom-up observation of the player's behaviors with little interpretation (e.g., observing repetitive behavior or no mistakes being made), while others involved top-down interpretation of the performance and the perception of ToM states (e.g., having goals during gameplay or experimenting with the game), or top-down interpretations that produced assumptions about the player without implying ToM states (e.g., not learning or having no goals while playing). Finally, one category was related to a pre-existing assumption that was expressed independently of observations of behavior (the belief that a skilled AI would make no mistakes). The overarching categories were related to whether the behavior being described was seen as salient and unpredictable (e.g., observing an agent make a mistake, which by nature are unintended and therefore typically not predictable, or perceiving that the player is learning or experimenting with the game, which can produce novel behavior), or

predictable or explainable in the sense that the behavior is simplistic and lacks remarkable actions (e.g., being repetitive or not making any mistakes), or that the behavior is viewed through an explainable lens when top-down interpretations are applied (e.g., performing actions to achieve an assumed goal or pausing for a moment to make a decision). Some categories were not exclusive to predictable or unpredictable behavior. Results including these mid-level and overarching categories are presented in Table 2 along with counts associated with perceived Humanness and Expertise.

Table 1: Summary of Coded Statements from Qualitative Analysis. The table shows the category, a representative example, the total count of coded statements associated with the category and relative frequency that it was mentioned by all participants, Cohen's Kappa as a measure of inter-rater reliability and a χ^2 between raters to show the likelihood of agreement between raters.

category	example	total count (and % of participants who mentioned code)	Cohen's Kappa	χ^2 between raters
AI makes no or few mistakes	"The Expert AI wouldn't make any mistakes"	17 (37%)	1	(1,324) = 302.11, p < 0.001
does not learn	"the player didn't adapt to the situation and made the same mistakes."	19 (52%)	0.721	(1,324) = 159.48, p < 0.001
experimenting	"Experimented with strategy. Seemed like it was learning"	41 (81%)	0.942	(1,324) = 278.5, p < 0.001
flowing	"AI because the agent knew how to time the jumps and didn't run into anything that would prevent it from continuing in the game."	16 (44%)	0.966	(1,324) = 281.91, p < 0.001
has goals or intentions	"Made an attempt to get more points and coins."	37 (63%)	0.969	(1,324) = 294.34, p < 0.001
has knowledge	"It knew the ways in which to complete the level."	29 (63%)	1	(1,324) = 311.85, p < 0.001
has no goals	"The agent just walked straight and didn't hit the boxes to collect coins."	8 (30%)	0.932	(1,324) = 242.73, p < 0.001
has no knowledge	"The agent would not jump pass the obstacle and didn't have knowledge of what to do next."	10 (33%)	1	(1,324) = 291.43, p < 0.001
has thoughts and reasoning	"The player had less mistakes and it second guessed in most cases."	14 (37%)	0.961	(1,324) = 276.26, p < 0.001
is skilled	"The agent did everything perfectly and there was no hesitation of anything"	19 (52%)	1	(1,324) = 286.3, p < 0.001
is unskilled	"The player was defeated more frequently and ran into a lot of enemies"	21 (48%)	0.974	(1,324) = 291.33, p < 0.001
learning	"Agent was taking time to learn strategy"	31 (59%)	0.924	(1,324) = 267.11, p < 0.001
makes decisions	"The Expert Human seemed to make human decisions and there was hesitation."	25 (56%)	0.797	(1,324) = 197.28, p < 0.001
makes mistakes	"Human because it would make mistakes similar to what I would do"	59 (85%)	0.957	(1,324) = 290.97, p < 0.001
no or few mistakes	"The agent did everything perfectly and there was no hesitation"	18 (44%)	0.869	(1,324) = 230.29, p < 0.001
precise	"very precise and methodical."	10 (26%)	1	(1,324) = 291.43, p < 0.001
repeatitive behavior	"Did the same thing every time"	44 (85%)	1	(1,324) = 315.54, p < 0.001
seeks information	"it wanted to make sure to play the game right because it read the instructions."	11 (37%)	0.951	(1,324) = 264, p < 0.001
unpredictable or random	"It wasn't perfect and it was going at different paces. The jumps were even different."	14 (44%)	1	(1,324) = 300.26, p < 0.001

Table 2: Categories Associated with Predictable and Unpredictable Behaviors and Counts for Perceived Humanness and Perceived Expertise. The higher count of perceived Humanness and Expertise is highlighted for each category.

			total count (and % of participants who mentioned code at least once)			
overarching category	**mid-level category**	**category**	**perceived as AI**	**perceived as H**	**perceived as B**	**perceived as E**
behavior is predictable or explainable	observation of base behavior without deeper interpretation	flowing	10 (30%)	6 (22%)	3 (7%)	13 (41%)
		is skilled	8 (30%)	11 (33%)	0 (0%)	19 (52%)
		no or few mistakes	11 (30%)	7 (22%)	1 (4%)	17 (41%)
		precise movements	7 (22%)	3 (11%)	2 (7%)	8 (22%)
		repeatitive behavior	32 (67%)	12 (33%)	31 (70%)	13 (37%)
	pre-existing assumption	AI makes no or few mistakes	4 (15%)	13 (30%)	6 (19%)	11 (30%)
	perception of ToM states	has goals or intentions	12 (26%)	25 (59%)	10 (37%)	27 (48%)
		has knowledge	11 (30%)	18 (44%)	7 (19%)	22 (56%)
		has thoughts and reasoning	3 (11%)	11 (30%)	7 (26%)	7 (22%)
		makes decisions	2 (7%)	23 (52%)	10 (33%)	15 (33%)
behavior is unpredictable	observation of base behavior without deeper interpretation	is unskilled	5 (19%)	16 (30%)	21 (48%)	0 (0%)
		makes mistakes	27 (59%)	32 (59%)	42 (81%)	17 (41%)
		unpredictable or random	2 (7%)	12 (37%)	11 (33%)	3 (11%)
	perception of ToM states	experimenting	4 (15%)	37 (70%)	22 (52%)	19 (56%)
		learning	3 (11%)	28 (56%)	23 (56%)	8 (22%)
		seeks information	3 (7%)	8 (30%)	10 (33%)	1 (4%)
describes either predictable or unpredictable behavior	assumptions based on perceived behavior/not ToM states	has no goals	4 (15%)	4 (15%)	8 (30%)	0 (0%)
		has no knowledge	7 (7%)	7 (26%)	7 (33%)	7 (0%)
		does not learn	17 (48%)	2 (4%)	18 (52%)	1 (4%)

Game Theoretic Analysis

Three typical game theory games were empirically derived from interview responses during the empirical game theoretic analysis: a coordination game, in which mutual competition was selected for fun, a Stag Hunt, in which mutual cooperation is selected for greater utility in achieving a higher score, and Deadlock, where the player chooses to compete regardless of what the other player decides because they believe they are more skilled and can win. Results can be seen in Table 3.

Table 3: Empirical Game Theoretic Analysis of Responses from the MarI/O Experiment. Three typical game theory games were derived from interview responses: a coordination game, in which mutual competition was selected for fun, a Stag Hunt, in which mutual cooperation is selected for greater utility in achieving a higher score, and Deadlock, where the player chooses to compete regardless of what the other player decides because they believe they are more skilled and can win.

Participant's Game:	Coordination	Stag Hunt	Deadlock
Example:	"I would play competitively against an Expert AI because it would be a challenge against a systematic agent. A beginner AI would be rather boring."	"I would play cooperatively with an Expert Human because the player would make sure to get the extra coins and would use Yoshi. We would get more points along the way"	"I would play competitively against a Beginner AI because it would be easier to beat the agent since it seldom knows what it's doing."
AI Beginner	0/27 (0%)	0/27 (0%)	11/27 (41%)
AI Expert	3/27 (11%)	8/27 (30%)	0/27 (0%)
Human Beginner	0/27 (0%)	0/27 (0%)	4/27 (15%)
Human Expert	8/27 (30%)	17/27 (63%)	0/27 (0%)

Discussion

From the coded interviews, a pattern began to emerge related to perceptions of human-likeness, competence, and behaviors that were described as random or unpredictable, or predictable and explainable in the sense that they emerged from having intentions and goals while playing the game.

In terms of the relationship between perceived human-likeness and competence, qualitative and quantitative data suggested there may be a non-linear relationship where

perceptions of human-likeness increased with agent competence but peaked at the level of intermediate.

The first indication of this non-linear relationship is in the quantitative data. Participants in this experiment perceived the beginner AI as a beginner and as an AI most frequently (rated beginner by 100.0%, SD = 0%, of participants in explicit ratings, and rated human by 22%, SD = 41.6%, of participants). On the opposite end of the spectrum, the expert human was rated as an expert most frequently (rated expert by 88.9%, SD = 31.4%, of participants), and rated as a human less frequently than the beginner human (74.1% and 84.0% rated human, respectively). Participants' explicit ratings of human-likeness and expertise can be seen in Figure 7. A second order polynomial trend line was added to this data which shows the peak perceived human-likeness at around 50% expertise, or intermediate level.

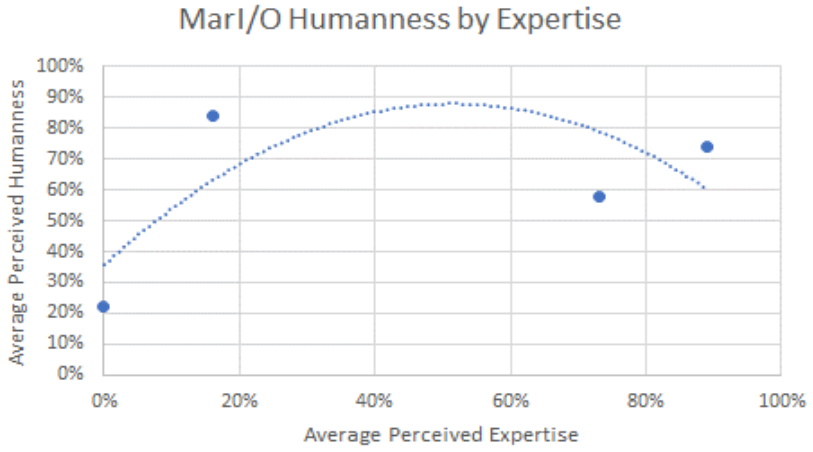

Figure 7: Explicit Ratings of Human-Likeness by Expertise in the MarI/O Experiment. A second order polynomial best fit line has been added, which shows that perceived human-likeness peaks at mid-level expertise.

The second indication of this non-linear relationship is related to a pre-existing assumption about highly expert or perfect performances that was stated by many of the participants. Many responses seemed to suggest that there may be a pre-existing assumption that at a competent level, humans still make mistakes while an AI would not (i.e., "AI makes no or few mistakes" category). Overall, 37%, SD = 48.2%, of participants explicitly stated this assumption. Some examples include:

> HE perceived as HE: "AI would have minimal mistakes."

> AE perceived as HE: "Human [because] the agent made a few errors. I don't think an expert computer would make those typical errors."

The expectation was generally used as a basis to compare observed behavior against in order to rate agents. This perception may be related to the perfect automation schema (Dzindolet et al., 2002), in which people have pre-existing assumptions that functional automation should have perfect performance.

The relationship between predictability and unpredictability also arose out of the analysis of the qualitative data. While analyzing the interviews, a number of concepts were coded from the cues participants used to distinguish between human and AI performances. Upon further analysis, the coded statements could be grouped by predictability or explainability, where statements related to given codes seemed to describe performances as either predictable/explainable or unpredictable/novel.

Table 2 shows the overarching categories, mid-level categories, and counts (with percentage of participants who had one or more codes in the category) associated with each category. A noticeable pattern emerged once the higher counts/percent for Human-likeness and Expertise were highlighted. Categories associated with unpredictability were more often associated with perceived human-likeness at a beginner level, whereas categories associated with predictability are often associated with either human-likeness at an expert level (i.e., competent performances), or AI-ness at a beginner level. The pattern suggests that perceived human-likeness and expertise may have a relationship with perceptions of behavior that is either seen as predictable or unpredictable.

When participants described performances, sometimes they described them in elaborate detail and referenced aspects of the player's cognition as they played through the game. For example:

> HB perceived as HB: *"Human because it knew how to stop to learn and think to figure out how to play the game. It wasn't perfect and it was going at different paces."*

> AE perceived as HE: *"The player was more selective and there seemed to be more strategy behind it." Or "The player did really well at the beginning, but had to think about which move to take while performing tasks."*

This may imply that they perceived an intelligent mind and ToM states in the player while watching them play the game that helped participants to understand and explain their actions in the game. Participants sometimes also described the performance in unpredictable terms, seeming to not know what the player would do next. For example:

> AE perceived as HE: *"More spinning than normally and didn't seem stuck in the same behaviors and it seemed like the player was having fun like a human would."*

> AB perceived as HB: *"Beginner Human [because] the player didn't collect coins and there was inconsistent and sporadic jumping."*

Furthermore, the agents' perceived human-likeness and competence/expertise seemed to interact with the overall perception of performances that were seen as predictable or unpredictable. Specifically, when a performance was perceived as predictable and skilled, descriptions seemed to indicate more perceived ToM states and more complex cognition (e.g., having goals and intentions, reasoning, making decisions, being precise) compared to predictable performances that were unskilled (e.g., repetitive, aimless). Some examples of predictable/explainable behavior include:

> AE perceived as HE: *"The expert human seemed to make human decisions and there was hesitation."*

AE perceived as HB: *"The player was still learning how to play, but they understood how to avoid losing the game and also to beat enemies"*

AB perceived as AB: *"The agent kept making the same patterns of mistakes."*

AB perceived as AB: *"The agent just walked straight and didn't hit the boxes to collect coins."*

As seen in these quotes, participants' language when describing these agents' behaviors sometimes include assumptions about the players' motivations and they remembered observing specific performances in the game. These quotes seem to speak to the observation of behaviors that are easily explainable, either due to being repetitive in nature or because the participant can intuitively understand the players' underlying goals or decisions.

When it came to descriptions of unpredictable behavior, even descriptions of unskilled behaviors that were unpredictable were often perceived as humanlike. Some examples include:

AB perceived as HB: *"Beginner Human [because] the player didn't collect coins and there was inconsistent and sporadic jumping."*

AB perceived as HB: *"The player was fairly slow and it didn't hit any targets. The player was ducking for no reason. The player would lose quite obviously to the enemy and it didn't know how to defend."*

AE perceived as HB: *"The player played longer, but there were more random tactics. It just raced through without a clear objective"*

While these participants described the players as having poor performance and perceived them as beginners, these AI players were also perceived as humans, which seems to be in part related to their unpredictable behavior.

The results from the empirical game theoretic analysis indicated that the strategic or social context perceived by participants can be easily mapped onto a few typical game theory games. While there may be some indication of an overall pattern (more complex contexts are perceived or preferred for competent humans, and simple competitive contexts are preferred for incompetent AI), there is not enough information available in this experiment to say anything conclusively.

Results from Experiment 1 suggest that the answer to research question 1 (How do people differentiate between actions produced by other humans and "mechanical" actions produced by AI?) is that participants may use pre-existing knowledge of how humans think and feel to interpret observations of behavior (i.e., ToM states). Additionally, participants may have pre-existing expectations of AI (i.e., the Perfect Automation Schema). Finally,

participants' perception of human-likeness may be influenced by how skilled or competent a player is perceived to be, as well as the presentation of novel and unpredictable behavior.

Overall, results from Experiment 1 suggest that two areas for follow-up research may be meaningful to explore: (1) the extent to which perceptions of skill/competence and predictability influenced the perception of human-likeness, and (2) how a more ecologically valid context of interaction may influence these overall perceptions, as well as how social relationships differ in Human-Human Interactions and Human-AI Interactions (HAI) after the interaction occurs.

Experiment 2, therefore, was developed to investigate skill/competence and predictability with more concrete, quantitative methods as opposed to the qualitative methods from Experiment 1.

Experiment 3 was developed to provide an immersive and complex social environment in which participants could make evaluations of human-likeness based on observations of complex behaviors (social and survival) while playing a videogame. Here, observable behaviors came from both the observation of how a co-player played the game and non-verbal social interaction, where participants and co-players (humans and AI) could interact however they desired within an open world survival videogame (Don't Starve Together).

EXPERIMENT 2: THE RELATIONSHIP BETWEEN PERCEIVED PREDICTABILITY, COMPETENCE AND HUMAN-LIKENESS IN MARI/O

To what extent do competence and predictability contribute to the perception of human-like behavior?

From Experiment 1, participants described agents in rich terms that were codified, revealing a complex relationship between perceived human-likeness, predictability, and competence. The second experiment is intended to investigate the extent to which competence and predictability contribute to the perception of human-like behavior in an attempt to answer the second research question: *to what extent do competence and predictability contribute to the perception of human-like behavior*? While the first experiment hinted at the type of relationship that may be present, Experiment 2 will answer the question in a more thorough and rigorous way and shed further light on the first research question.

While competence was intentionally varied for all experiments, predictability was not. In Experiment 1, participants described behaviors in terms that either indicated that they perceived different agents' behaviors as predictable and explainable (e.g., repeating an action or acting in a way that was intuitive if the player had human-like motivations and mind) or as unpredictable or random (e.g., being aimless, jumping at random times, or experimenting with the environment in new ways). While seeing an agent as predictable can be associate with very simple behaviors like repeatedly performing the same action, a humanlike mind can also be considered predictable when we can intuit the goals and intentions (i.e., Theory of Mind; Premack & Woodruff, 1978; Baron-Cohen, Leslie & Frith,

1985). Similarly, it is possible that participants perceived these actions as explainable in that they resulted from trying to achieve goals, and the perception of goal-directedness is associated with the perception of animate motion (Rakison & Poulin-Dubois, 2001). In Experiment 1, participants who perceived an AI agent as competent and described them in predictable terms often perceived the agent as a human. In addition, these terms were often associated with understanding the cognition and intentions of the agent. For example:

> AE perceived as HE: "Human because there was some hesitation and selective human behavior."

While the participant could have perceived pauses in the agent's motion in a variety of ways, it is described as "hesitation" and "selective", as if to imply that the agent was making purposeful decisions while playing the game.

Additionally, acting predictability in a joint action task can also lead to the perception that an interaction partner is trying to coordinate with their interaction partner, which can also imply human-likeness. Pfeiffer et al. (2011) demonstrated that in a naive context, the act of frequently engaging in joint attention (i.e., predictably looking at the same object that a participant was looking at) increased perception of human-likeness, whereas in a cooperative context (i.e., trying to engage in joint attention as often as possible), the perception of human-likeness was increased both when joint attention was frequently engaged, as well as when it was frequently avoided (i.e., predictably looking away for the object a participant was looking at). This indicates that the perception of predictability can influence the perception of human-likeness, and that this perception is dependent on the context of an interaction. In the same study, a competitive context (i.e.,

trying to avoid joint attention) eliminated the correlation of predictability and perception of human-likeness. This may imply that overall "predictability" is not just perceived in relation to observed behavior but is accompanied by the interpretation of the intentions behind that behavior. It is possible that participants in the Pfeiffer study who were in a cooperative context but had an interaction partner that never successfully engaged in joint attention believed their interaction partner was trying to engage, and was just really bad at the task, or that they were intentionally toying with them, whereas in a competitive context, the interpretation of intentions may not have been as clear. The ability to perceive intentions in the interaction partner may have increased the perception of human-likeness.

On the other end of the spectrum, the observation of random or unpredictable behavior has been observed to increase with anthropomorphism. One relatively easy method for designing behavior that is perceived as human-like is to have a robot or AI agent act in ways that appear random, unpredictable, or violate people's expectations, although this can also cause these agents to be perceived as less attractive or likable (Waytz et al., 2010b; Short et al., 2010; Hayes et al., 2014). There are two note-worthy exceptions in recent studies (Salem et al., 2013; Kompatsari et al. 2019). When performing an unpacking task, ASIMO, the large, white, humanoid robot built by Honda (Shigemi et al., 2019), was observed occasionally making incorrect hand gestures (incongruent direction as a verbal instruction). These "errors" increased anthropomorphism, the extent to which participants liked the robot, and interest in future interactions even while participants knowingly performed worse at the task (Salem et al., 2013). Kompatsari et al. (2019)

demonstrated that a robot which engaged in joint eye gaze independent of task increased ratings of human-likeness and likability. However, both studies involved a technologically sophisticated, embodied robot in a non-competitive task (compared to the competitive tasks in Short et al., 2010 and Hayes et al., 2014). So factors like amusement and novelty may have played a role during the studies but might not be present if such agents were a normal part of life. Of additional consideration is that different modes of unpredictable behavior (for instance, the verbal cheat and active cheat conditions of Short et al., 2010) can affect mind perception in dramatically different ways (i.e., "it malfunctioned", or "it tried to cheat", respectively), so it is important to consider how different behaviors communicating the same intent can lead to distinct interpretations. While a simple set of behavioral characteristics that can effectively trigger anthropomorphism and mind perception would be desirable for robot and AI designers (Wiese, Metta & Wykowska, 2017), more work is needed to elucidate the types of perceived behaviors lead to desired interpretations in HAI and HRI.

Hypotheses for Experiment 2

The hypotheses for this experiment are that (H1) the relationship between competence and perceived human-likeness will be non-linear, and that as perceived competence increases, perceived human-likeness will generally increase, but will peak at the level of intermediate skill. Perception of human-likeness will decrease at the level of expertise due to perceptions related to the Perfect Automation Schema (Dzindolet et al., 2002), which was seen in the results from Experiments 1 and 2. An example of what this

may look like is presented in Figure 8. Furthermore, (H2) the relationship between predictability/explainability and perceived human-likeness will be affected by perceived competence, which is presented in Figure 9. At a beginner level, highly predictable behaviors will be perceived with low human-likeness, and human-likeness will generally increase as unpredictability increases because predictability at a beginner level is likely to be associated with repetitive behavior rather than complex but highly explainable behaviors, and unpredictability has been shown to increase anthropomorphism (Waytz et al., 2010b; Short et al., 2010; Hayes et al., 2014). Intermediates will have a nonlinear relationship, in which performances perceived as highly predictable/explainable or highly unpredictable will be perceived with high human-likeness because the predictable behaviors will be explainable as ToM states (e.g., reasoning, having knowledge), as was seen in results from Experiment 1, and unpredictability will again increase anthropomorphism. The pattern for experts may be somewhere between these two, as behaviors will be complex enough to demonstrate complex mental states, but the perception of human-likeness will be muted overall due to the pervasive opinion that expert level performances are more likely to come from AI.

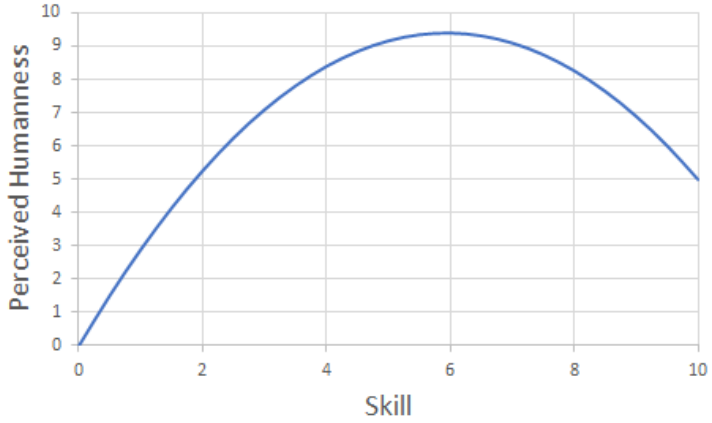

Figure 8: Hypothesized Human-Likeness and Skill Results from Experiment 2. This corresponds with H1. It is predicted that perceived human-likeness will in general increase with perceived skill, but that the peak value will be at the level of intermediate.

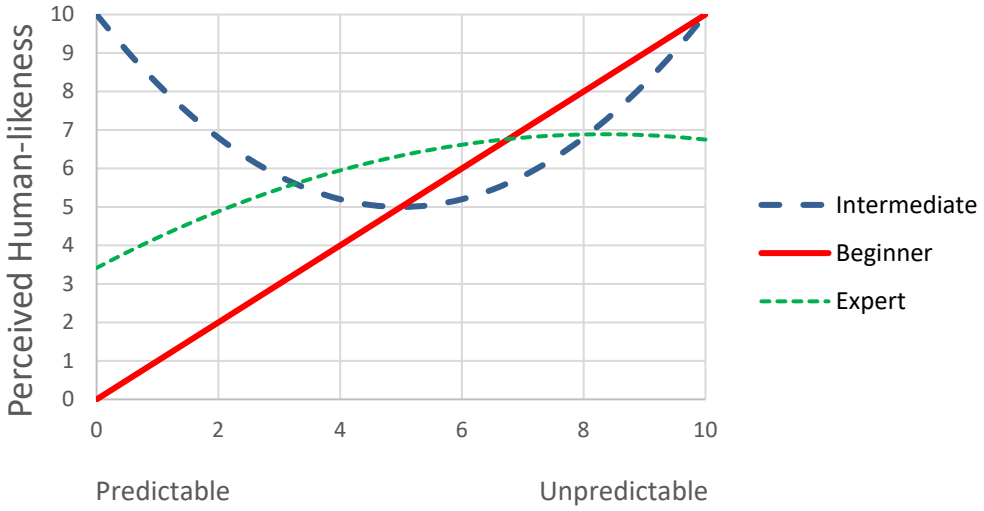

Figure 9: Hypothesized Perceived Human-Likeness and Perceived Predictability/Unpredictability by Expertise Results from Experiment 2. This corresponds with H2. It is predicted that agents perceived as beginners, intermediates and experts will have different relationships between perceived human-likeness and predictability, with a positive linear relationship for beginners, a u-shaped relationship for intermediates, and something between the two for experts.

Participants

The experiment includes 111 participants, where 100 were from Amazon's Mechanical Turk (mTurk; mean age: 36.43, SD: 10.06; 48 females) and 11 from George Mason's Sona System (mean age: 25.45, SD: 5.57; 11 females). The screening criteria were that

participants be over 18 years of age and currently reside in the US. In total, 122 participants were collected from mTurk, but 22 participants were removed from the data set due to failing an attention check question (see Measures), finishing faster than possible (faster than 17 minutes), answering 90% or more of trials in the exact same way, or having an IP address that indicated that they resided outside of the US. No participants were removed from the Sona sample.

A post hoc power analysis was conducted, assuming a medium effect size $f^2 = .15$, α error probability of 0.05 and 2 predictors (skill and predictability), a sample size of 111 yielded a power of 0.96.

Stimuli

The stimuli in this experiment included vignettes of game play from the same AI and human players as Experiment 1 playing the first level of Super Mario World. The one exception was that the AI beginner (AB), which was replaced with a version with slightly more training than the AB used in Experiment 1. The purpose of this substitution was that the AB in Experiment 1 was labeled as an expert much less often than the HB and many of the performances were repetitive, which likely affected overall perceptions of human-likeness and skill.

There were 6 vignettes for each player. All vignettes were between 3 and 8 second in length. Each vignette was intended to showcase one type of behavior at a time (e.g., interacting with enemies, having a pattern of jumping, collecting coins, or pausing to consider a decision). There were 6 vignettes for each player. Each set of vignettes were

balanced such that both beginners had 2 vignettes displaying an obvious mistake (dying after being hit by an enemy) and experts had 1 vignette with an obvious mistake. All player vignettes featured 1 vignette where the player completed the level, and all players had vignettes featuring similar experiences from throughout the level. Vignettes were converted into GIFs that played on a loop with approximately 500ms of black screen at the beginning of the vignette to indicate when it began again.

Apparatus

Participants used their own computers to complete the study. The experiment was hosted on Qualtrics and completed online in one sitting. Participants found the study either through signing up on the Sona System website or through Amazon's Mechanical Turk (mTurk).

Measures

All measures for this study were quantitative. Participants responded to each question via mouse click. All measures were recorded via Qualtrics.

For each trial, participants watched a brief vignette, then indicated their perceptions of how likely a player was human, the player's skill level at the game, and how predictable they believe the player's actions were on 10 point Likert scales. The questions read as follows:

1. How likely is this agent a human?
2. How would you rate the skill of this agent?

3. How predictable were this agent's actions?

After responding to the three Likert measures, participants were asked to select character traits they associated with the player given the performance they just watched. The question and 18 character trait options are presented in Table 4. These traits were presented in 3 pages with 6 on each page. These traits come from the categories derived from qualitative data in Experiment 1. Some traits are presented with their opposite (e.g., "makes mistakes" and "makes no mistakes"). Whenever an opposite occurred, the pair of traits was presented side by side on a page.

Table 4: Character Trait Question and Character Traits Derived from Experiment 1. In Experiment 3, participants watched a vignette and selected any and all traits they associated with the player. Traits were presented in 3 groups (seen in columns) so that participants were not overwhelmed when making selections. These were always presented in the same order to make it easier to respond over the 24 trials. The same presentation of terms occurred for the last 2 questions about character traits typically associated with humans and AI.

Question to participants: Please select any and all terms you would associate with this agent		
Options		
makes mistakes	thinking	experiments with game
makes no mistakes	flowing	random
makes decisions	learns from actions	has knowledge
repetitive	doesn't learn	no knowledge
precise actions	has goals	has skills
seeks information	has no goals	has no skills

After each block, participants were asked about their overall perceptions of human-likeness, expertise and predictability on 10-point Likert scales, and about their strategic preferences for playing a game with the player in the future, and asked to indicate a reason. The options given have strategies and justifications that match the game theoretic games identified in Experiments 1 & 3. It can be noted that Experiment 3 was conducted prior to Experiment 2, so the insights were used. The question, options, and associated game theory games are presented in Table 5.

Table 5: Strategic Preference Question, Options, and Corresponding Game Theoretic Games. Participants are asked about their strategic preferences after each block and asked to indicate their preference along with a justification. Each preference and justification corresponded to a specific game theoretic game.

Question to participants: If you were to play a game in the future with this agent, would you prefer to play cooperatively or competitively, and why?	
Option	**Game Theoretic Game**
Cooperatively, because I prefer cooperative games	Coordination, with cooperation as preference
Cooperatively, because we could do better working together	Stag Hunt
Cooperatively, because I don't want to lose	Mutual Assured Destruction
Competitively, because it's more fun	Coordination, with competition as preference
Competitively, because I think I would win easily	Deadlock
Competitively, because I couldn't trust that they would cooperate	Social Dilemma

At the end of all four blocks, participants selected any and all character traits they typically associated with humans and AI. The questions read as follows (terms seen in Table 5):

1. Please select any and all terms you would typically associate with human players.

2. Please select any and all terms you would typically associate with AI players.

On the version that is placed on mTurk, two attention check questions were included and presented at random points in the experiment after a block within the Likert questions. The questions read as follows:

1. If you are paying attention right now, select three.
2. If you are paying attention right now, select eight.

Procedure

Participants first gave informed consent and answered a brief demographic questionnaire. Participants were instructed that they would be observing the 4 players (HB,HE,AB,AE) play Super Mario World and evaluating the players based on their (the participant's) overall perceptions (for a detailed description of the game, refer to Experiment 1).

Next participants were given instructions and information on the structure of the experiment. Throughout the experiment, participants watched vignettes, then rated how likely the player was human, how skilled the player was, and how predictable their actions were based on the performance through Likerts (see Measures). After the Likerts, participants selected any and all of the 18 character traits they associated with that player based on the vignette. Since the vignettes were presented as GIFs that repeated, they could watch the GIF repeat as much as they desired.

Participants were informed that they would watch each of the 4 agents in 4 individual blocks, where each block included 6 trials from one of the agents. Participants were only informed that some of the players were humans and some were AI, and not that

we classified some as beginners and experts. All blocks were presented in random order, and all trials were randomized within blocks.

Before beginning the experiment, participants watched a video of the entire first level being played through so that they knew what to expect throughout the trials. After reading the instructions and watching the video, participants completed one practice trial. The vignette played in the practice trial and the video of the first level did not include any footage that was included in the experimental stimuli (i.e., the vignettes for each player).

At the end of the experiment, participants selected any and all of the 18 character traits they typically associated with human and AI players (two separate questions, see Measures). These two questions were presented in random order to participants. The entire experiment took approximately 30 – 45 minutes, depending on the participant's pace.

Quantitative Analysis

First, all measures based on 1 to 10 Likert scales were transformed by subtracting 1, such the lowest values were 0 and the highest values were 9. Analysis of Variance (ANOVA) was used to understand how players were rated differently across the three measures: perceived human-likeness, perceived skill, and perceived predictability.

A linear regression was conducted to understand how average perceived skill affected average perceived human-likeness. Next, a second order polynomial contrast was included in the regression of average perceived skill on average perceived human-likeness, and the two models were compared with an ANOVA to determine if one was significantly better at explaining the variance.

A multiple regression with a polynomial contrast were conducted to determine the effect of average perceived predictability on average perceived human-likeness for 3 levels of average perceived skill: beginner, intermediate, and expert.

Next, a cluster analysis was conducted to understand the relationship between the character traits (i.e., derived categories from Experiment 1) and the three measures: perceived human-likeness, perceived skill, and perceived predictability. The same cluster analysis was then repeated to understand the relationship between the Game Theoretic games (GT games). Additionally, a chi-squared analysis was used to investigate the relationship between agents and GT games, then ANVOAs were used to understand any relationship between GT games and average perceived human-likeness, skill and predictability.

Results

Perceptions of Individual Players

To gain a better understanding of how each player was perceived, 3 ANOVAs were conducted for average perceived human-likeness, average perceived skill, and average perceived predictability with respect to players. Average perceived human-likeness was not significantly different across players, $F(3,440) = 1.63$, $p = 0.181$, $\eta_p^2 = 1.011$, nor was average perceived predictability across players, $F(3,440) = 0.50$, $p = 0.685$, $\eta_p^2 = 0.003$. Average perceived skill was significantly different across players, $F(3,440) = 7.40$, $p < 0.001$, $\eta_p^2 = 0.048$. A post hoc Tukey's HSD showed that there were significant differences between the AE (mean = 5.74, SD = 2.46) and AB (mean = 4.86, SD = 2.70; $p = 0.002$),

the HE (mean = 5.56, SD = 2.61) and AB (p = 0.025), the HB (mean= 4.82, SD = 2.70) and AE (p = 0.001), and the HE and HB (p = 1.015). The differences were not significant between the HB and AB (p = 0.998) nor the HE and AE (p = 0.881). Results can be seen in Figure 10.

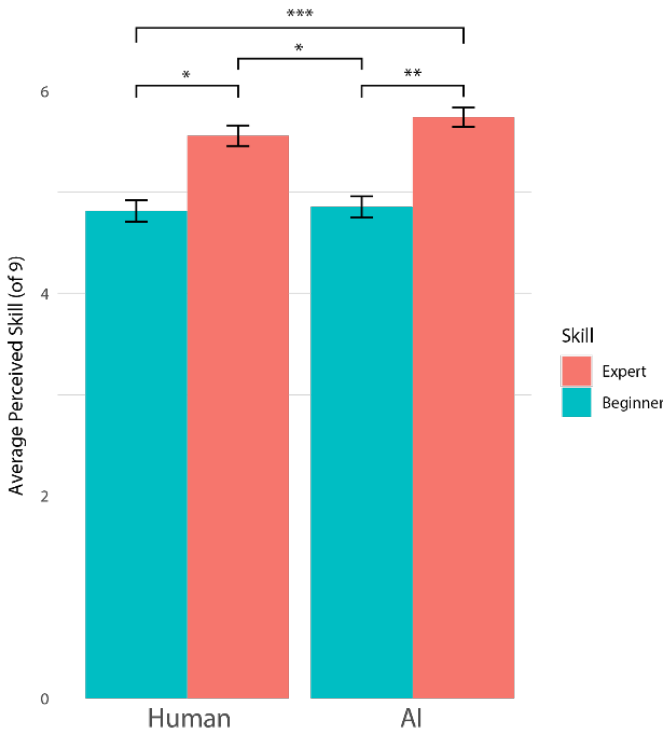

Figure 10: Perceived Skill by Player. There were no significant differences between the beginner players, or the expert players, but all other differences were significant. Standard error bars are presented with average perceived skill.

Perceived Human-likeness and Skill

Hypothesis 1 was that the relationship between perceived human-likeness and perceived skill would by polynomial in nature, with a general increase in perceived human-likeness as perceived skill increased, and higher levels of perceived human-likeness at the intermediate level. To examine this, first a linear regression was run for average perceived

human-likeness by average perceived skill. The model was significant, $F(1,442)= 129.70$, $p < .001$, $R^2 = 0.23$, and the equation is average perceived human-likeness = average perceived skill * 0.038 + 3.53. Next, a regression with a polynomial contrast was run for average perceived human-likeness by average perceived skill. The model was significant, $F(2,441)= 74.77$, $p < .001$, $R^2 = 0.25$, where R^2 was higher than in the first model, indicating that it explained more of the variance. The equation is average perceived human-likeness = (average perceived skill)^2 * 0.064 + 5.00. Next, the two models were compared with an ANOVA, which showed that the second model which included the polynomial contrast did lead to a significantly better model fit than did the linear model, $F(1,441) = 15.53$, $p < 0.001$, which confirms part of H1.

However, when looking at the equation and a scatter plot with the best fit line (polynomial, see Figure 11), the direction of the polynomial trend line was different than hypothesized in that perceived human-likeness was lower at the perceived level of intermediate. Upon inspection of the scatterplot, it was observed that there was a high degree of variability for players rated below the expert level which appeared to be driven by the variability in perceived human-likeness, where players with a high average perceived human-likeness had wide variability in perceived skill (seen in Figure 11). Since the second hypothesis necessitated the observation of low, medium and high levels of perceived skill in relation to perceived human-likeness and perceived predictability, a post hoc analysis of low, medium and high levels of perceived human-likeness in relation to perceived skill and perceived predictability was also examined.

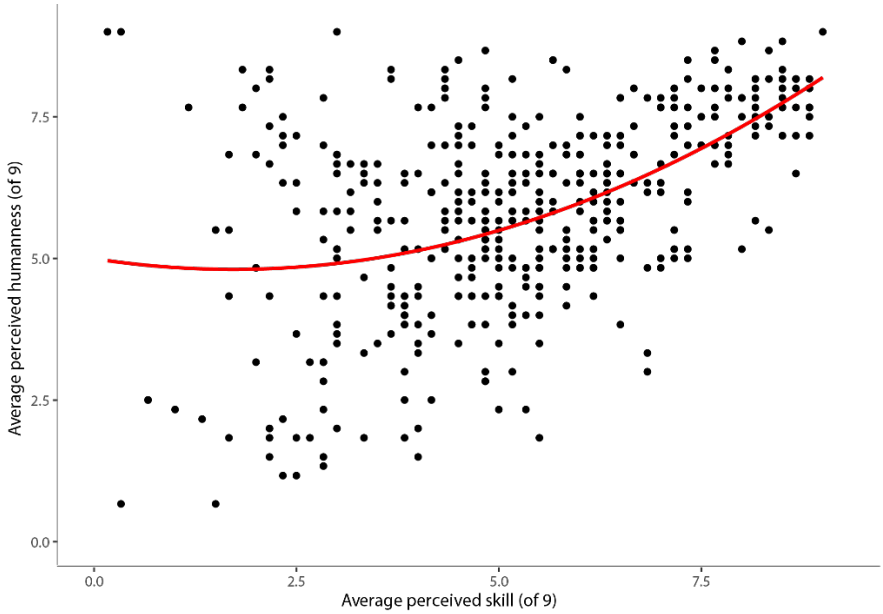

Figure 11: Average Perceived Human-Likeness by Average Perceived Skill. The best fit trendline has been added to the figure, and is non-linear. It was determined that the trendline is quadratic in nature.

Perceived Human-likeness in Relation to Perceived Predictability and Skill

Hypothesis 2 was that there would be differing patterns for perceived human-likeness and perceived predictability according to 3 levels of perceived skill: beginner, intermediate, and expert. In particular, it was hypothesized that there would be a linear pattern at the level of beginner, where an increase in perceived unpredictability would be associated with an increase in perceived human-likeness, and that a polynomial trend would exist for intermediate and expert levels, where perceived human-likeness would be high at the low and high ends of perceived predictability; there was not a strong hypothesis for experts, but it was believed that the pattern would be between that of beginner and intermediate, which was directly related to H1, where it was hypothesized that perceived human-likeness would be lower for experts than intermediates.

To assess H2, again, a linear regression was first run for average perceived human-likeness by average perceived predictability. The model was significant, $F(2,441)=90.79$, $p < .001$, $R^2 = 0.17$, and the equation is average perceived human-likeness = average perceived predictability * 0.38 + 3.81. Next, a regression with a polynomial contrast was run for average perceived human-likeness by average perceived skill. The model was significant, $F(2,441)=58.24$, $p < .001$, $R^2 = 0.21$, where R^2 was higher than in the first model, indicating that it explained more of the variance. The equation is average perceived human-likeness = average perceived predictability * -0.47 + (average perceived predictability)^2 * 0.08 + 5.74. Next, the two models were compared with an ANOVA, which showed that the second model which included the polynomial contrast did lead to a significantly better model fit than did the linear model, $F(1,441) = 21.48$, $p < 0.001$.

Next, a multiple regression with a polynomial contrast was run for average perceived human-likeness with respect to average perceived predictability with 3 levels of average perceived skill: beginner (average rating of 3 or less, n = 154), intermediate (average rating of 7 or less, n = 148), and expert (average rating above 7, n = 142). These groups are as close to balanced as could be achieved. The multiple polynomial regression equation was significant, $F(8,438) = 18.44$, $p < 0.001$, $R^2 = 0.24$.

An ANOVA was used to compare the regression to a model that did not include levels of average perceived skill. While that equation was also significant, $F(2,441) = 58.24$, $p < 0.001$, $R^2 = 0.21$, the ANOVA revealed that the equation that included levels of average perceived skill was a significantly better fit, $F(6,435) = 4.30$, $p < 0.001$.

Since average perceived skill was a categorical variable, an ANOVA was run on the regression model of average perceived human-likeness with respect to average perceived predictability with 3 levels of average perceived skill to compare the main effects and interaction. The main effect of average perceived skill level was significant, $F(2,435) = 8.93$, $p < 0.001$, $\eta_p^2 = 0.039$, as were the linear trend of average perceived predictability, $F(2,435) = 99.26$, $p < 0.001$, $\eta_p^2 < 0.001$, and the quadratic trend, $F(2,435) = 22.45$, $p < 0.001$, $\eta_p^2 = 0.009$. The interactions between the linear trend of average perceived predictability and average perceived skill level was not significant, $F(2,435) = 1.41$, $p = 0.244$, $\eta_p^2 < 0.001$, nor was the interaction between the quadratic trend and average perceived skill level, $F(2,435) = 2.55$, $p = 0.079$, $\eta_p^2 < 0.001$, though it was close to significant. A scatterplot with best fit regression lines for each level of skill is presented in Figure 12.

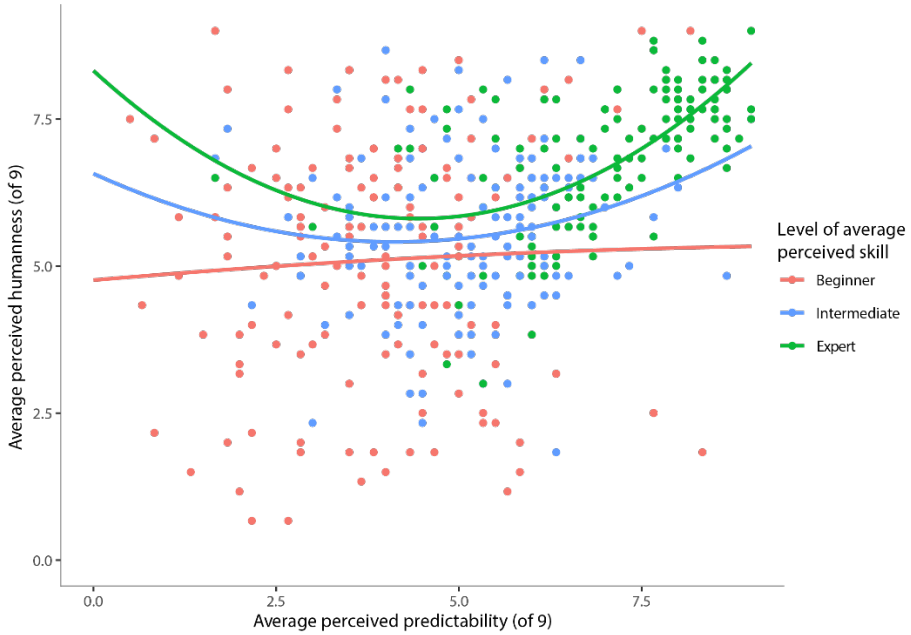

Figure 12: Average Perceived Human-Likeness by Average Perceived Unpredictability. Predictability is reverse coded to give unpredictability, which is what was presented with H2. Regression lines are presented for average perceived skill levels of beginner, intermediate, and expert. There was a main effect of average perceived skill level, and linear and quadratic trends for perceived predictability. It can be seen that at the level of expert, low and high predictability

are related to higher average perceptions of human-likeness. The same type of non-linear pattern is present for the levels of intermediate and beginner, but average perceived human-likeness is lower for lower levels of perceived skill.

As a post hoc analysis, since average perceived human-likeness seemed to drive the variability in Figure 11, regressions were run for average perceived predictability by average perceived skill. First, a linear regression was conducted to see, which was significant, $F(1,442) = 386.5$, $p < 0.001$, $R^2 = .47$. Next, a regression was run with a second order polynomial contrast, which was also significant, $F(2,441) = 236.10$, $p < 0.001$, $R^2 = .51$, and explained more of the variance than the linear model. Next an ANVOA was run on the two models, and it was determined that the equation with the polynomial contrast was significantly better at explaining the variance than was the linear model. The equation is average perceived predictability = $F(1, 441) = 46.23$, $p < 0.001$.

Next, a multiple regression with polynomial contrasts was run for average perceived predictability with respect to average perceived skill with 3 levels of average perceived human-likeness: low (average rating of 5.2 or less, n = 156), medium (average rating of 6.7 or less, n = 141), and high (average rating above 6.7, n = 147). These groups are as close to balanced as could be achieved. The multiple polynomial regression equation was significant, $F(8,435) = 62.81$, $p < 0.001$, $R^2 = 0.53$.

An ANOVA was used to compare the results of this model to the model without the 3 levels of average perceived human-likeness, and it was determined that the equation that included levels of average perceived human-likeness was a significantly better fit, $F(6,435) = 2.96$, $p = 0.008$.

Since average perceived human-likeness was a categorical variable, an ANOVA was run on the regression model to compare the main effects and interaction. The linear trend of average perceived skill was significant, $F(2,435) = 437.34$, $p < 0.001$, $\eta_p^2 < 0.001$, and quadratic trend, $F(2,435) = 47.45$, $p < 0.001$, $\eta_p^2 = 0.042$. The main effect of average perceived human-likeness level was not significant, $F(2,435) = 2.70$, $p = 0.068$, $\eta_p^2 = 0.012$, though it was close to significant. The interaction between the linear trend of average perceived skill and average perceived human-likeness level was not significant, $F(2,435) = 1.33$, $p = 0.266$, $\eta_p^2 < 0.001$, but the interaction between the quadratic trend and average perceived human-likeness level was significant, $F(2,435) = 4.81$, $p = 0.009$, $\eta_p^2 = 0.022$. Results are presented in Figure 13.

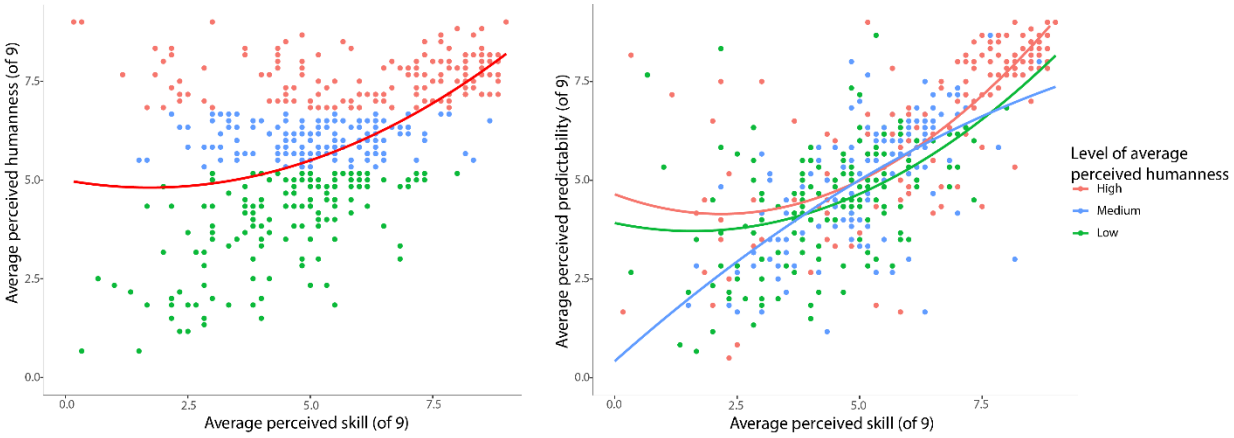

Figure 13: Average Perceived Predictability for Levels of Perceived Human-Likeness. (left) Average perceived human-likeness by average perceived skill. (right) Average perceived predictability by average perceived skill and level of perceived human-likeness. The regression lines are included for high, medium and low levels of average perceived human-likeness.

Finally, since the main effect of average perceived human-likeness level was nearly significant, the levels were next split as follows: low (average rating of 3 or less, n = 35),

medium (average rating above 3 and up to 6, not including 3, n =191), and high (average rating above 6, n = 218). When the polynomial multiple regression was run for perceived predictability by perceived skill with these 3 levels of perceived human-likeness, the equation was significant, $F(8,435) = 65.72$, $p < 0.001$, $R^2 = 0.54$. The ANOVA conducted on the regression model showed a significant main effect of average perceived human-likeness level, $F(2,435) = 6.02$, $p = 0.002$, $\eta_p^2 = 0.027$, as well as the linear trend of average perceived skill, $F(2,435) = 448.22$, $p < 0.001$, $\eta_p^2 < 0.001$, and the quadratic trend of average perceived skill, $F(2,435) = 48.63$, $p < 0.001$, $\eta_p^2 = 0.038$, and the interaction between levels of average perceived human-likeness and the quadratic trend of average perceived skill, $F(2,435) = 6.20$, $p = 0.002$, $\eta_p^2 = 0.028$. The interaction between the level of average perceived human-likeness and the linear trend of average perceived skill was not significant, $F(2,435) = 2.25$, $p = 0.107$, $\eta_p^2 < 0.001$. Figure 14 shows a comparison of the results presented in Figure 13 with balanced group sizes to these results with unbalanced group sizes at even levels. It is important to note that this model (with unbalanced groups) explained more of the variability than did the prior model (with balanced groups; $R^2 = 0.54$ and $R^2 = 0.53$, respectively), though an ANOVA could not be run on the two models because the model was saturated (same variables, only slightly different groupings), so there were no remaining degrees of freedom to run statistics on.

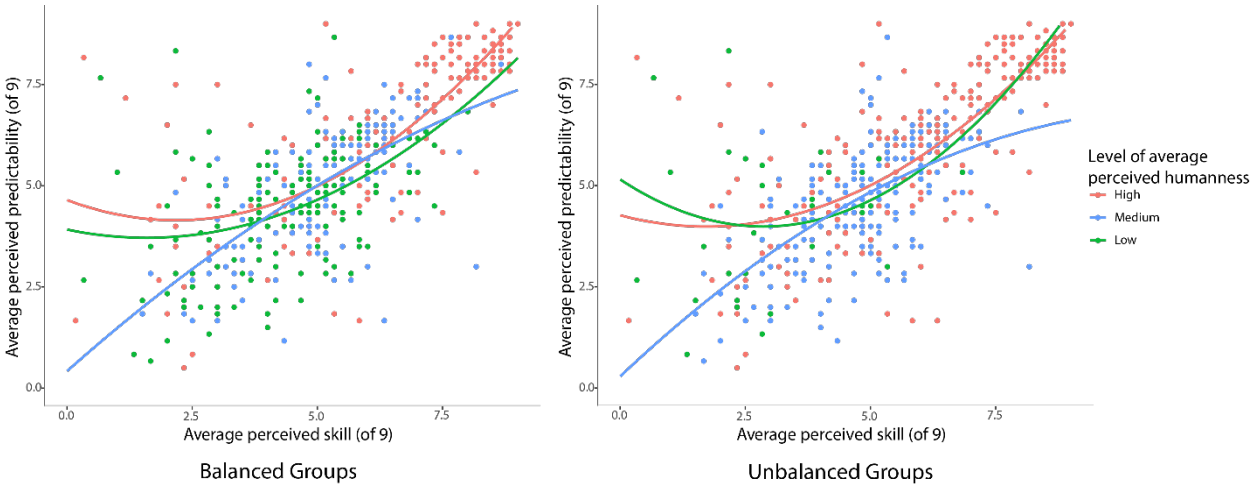

Figure 14: Average Perceived Predictability by Average Perceived Skill Depending on Levels Of Humanness and Group Sizes. (left) Average perceived predictability by average perceived skill and level of perceived human-likeness with balanced groups. (right) Average perceived predictability by average perceived skill and level of perceived human-likeness with unbalanced groups and even bin sizes. While a significant main effect of level of human-likeness was achieved with even bin sizes (nearly significant with balanced group sizes), the regression lines in each figure are very similar, indicating that the same basic pattern exists between the two level distinctions.

Cluster Analysis of Character Traits.

Character traits selected by participants were analyzed both in terms of the traits participants selected for each trail and the overall character traits participants typically associated with humans and AI. A cluster analysis could not be performed for the overall traits typically associated with humans and AI because these final questions were not accompanied by any trials or the three measures (perceived human-likeness, skill and predictability). Instead, the overall percent of participants who typically associated traits with humans and AI was examined, and these results can be seen in Figure 15. These character traits were sorted by the difference between percentages (i.e., percent AI – percent H) to show which traits are more commonly associated with humans and AI.

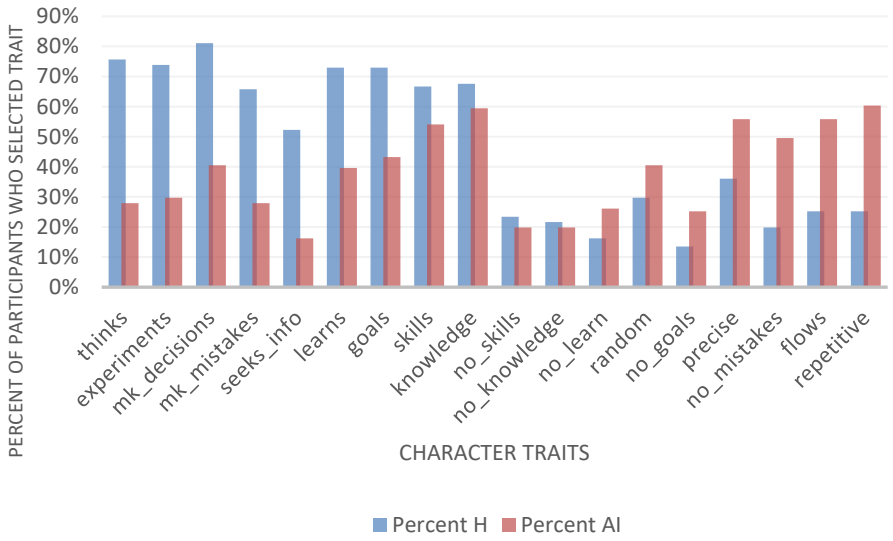

Figure 15: Percent of Participants who Typically Associated Each Character Trait with Humans and AI. These percentages are sorted such that traits more associated with human-likeness (i.e., difference between two percentages is high with a higher overall percent for humans) are on the left and traits more associated with AI-ness (i.e., difference between two percentages is high with a higher overall percent for AI) on the right. Percentages that are close to equal are in the middle.

A cluster analysis was conducted for the character traits with respect to the averages of perceived human-likeness, skill and predictability. To do this, for every character trait, every trial in which that character trait was selected was marked, and the perceived human-likeness, skill and predictability participants indicated for that trial was included in an overall average of the three measures for that character trait.

First, the Euclidean equation was used to calculate the correlation distance between measures (see Figure 16).

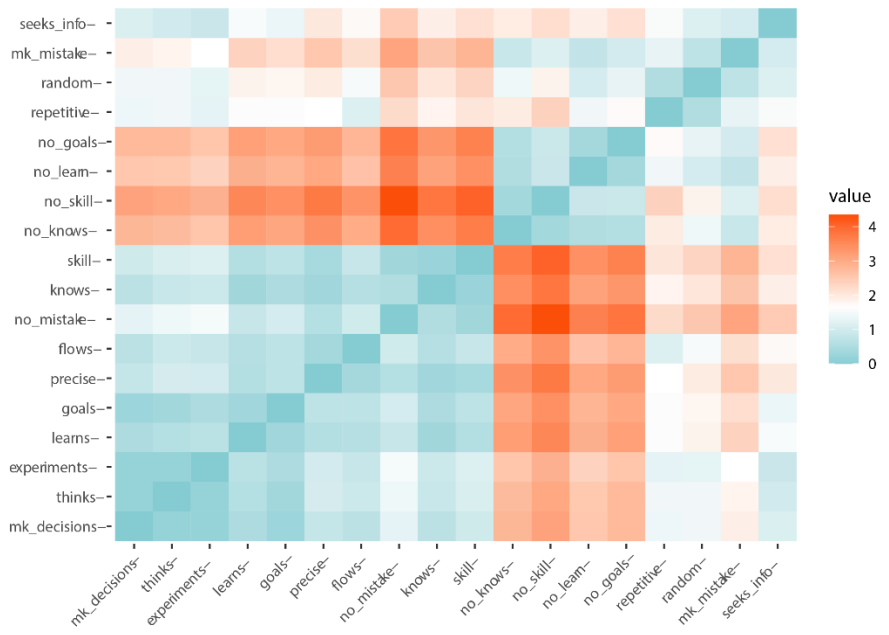

Figure 16: Results from Euclidian Distance Equation for Character Traits. Higher values (in red) indicate larger distance between terms.

Next, the Elbow method was used to estimate the optimal number of clusters, and 2 was determined to be the best number (see Figure 17).

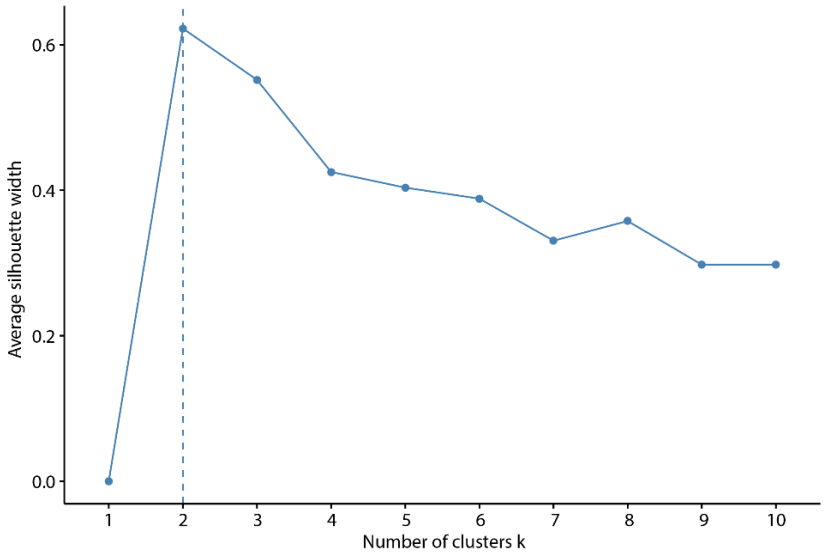

Figure 17: Optimal Number of Clusters According to Elbow Analysis for Character Traits.

Next, a k-means clustering unsupervised learning algorithm was used to separate the character traits into groups (see Figure 18). The average perceived human-likeness is 6.35 for cluster 1 and 7.02 for cluster 2. Average perceived skill is 4.72 for cluster 1 and 6.93 for cluster 2. Average perceived predictability is 5.76 for cluster 1 and 6.52 for cluster 2. ANOVAs demonstrated that average perceived human-likeness by cluster was significant, $F(1,16) = 13.9$, $p = 0.002$, part eta = 0.465, average perceived skill by cluster was significant, $F(1,16) = 65.02$, $p < 0.001$, part eta = 0.803, and average perceived predictability by cluster was significant, $F(1,16) = 19.50$, $p < 0.001$, part eta = 0.549.

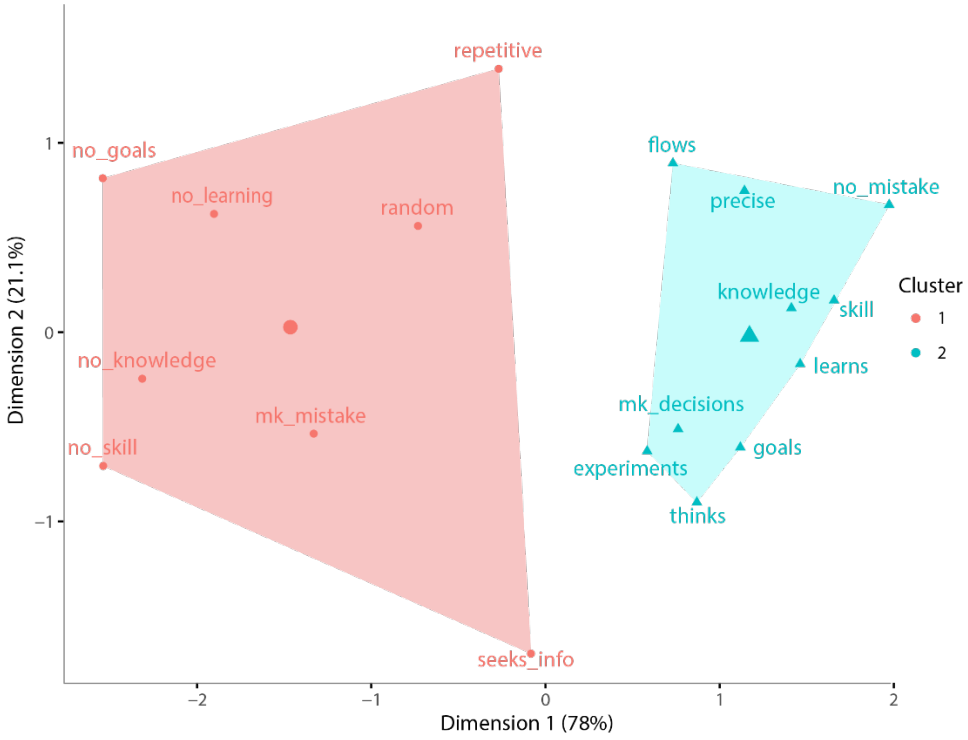

Figure 18: Cluster Plot of Character Traits.

Finally, individual plots were created for character traits along the dimensions of average perceived human-likeness and average perceived skill (Figure 19) and average perceived human-likeness and average perceived predictability (Figure 20).

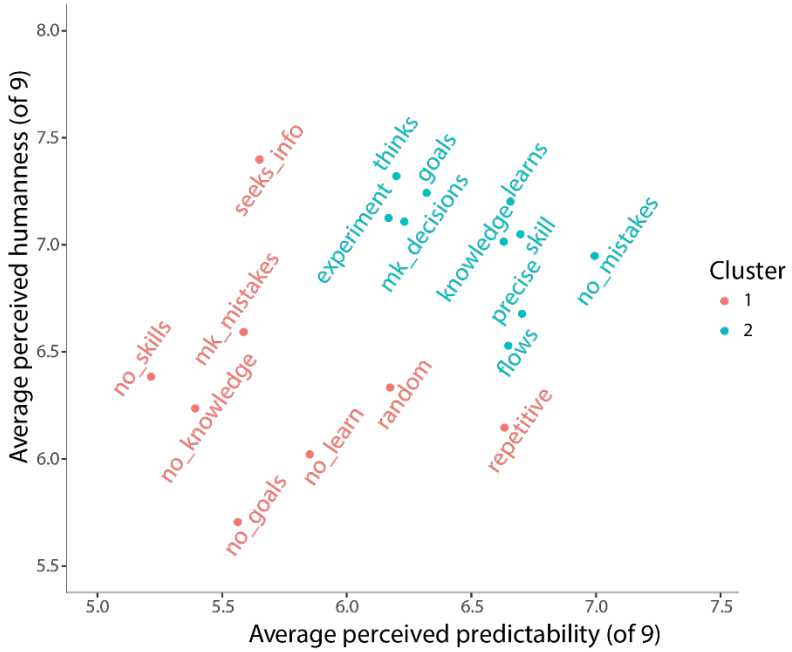

Figure 19: Character Traits by Average Perceived Human-Likeness and Average Perceived Skill.

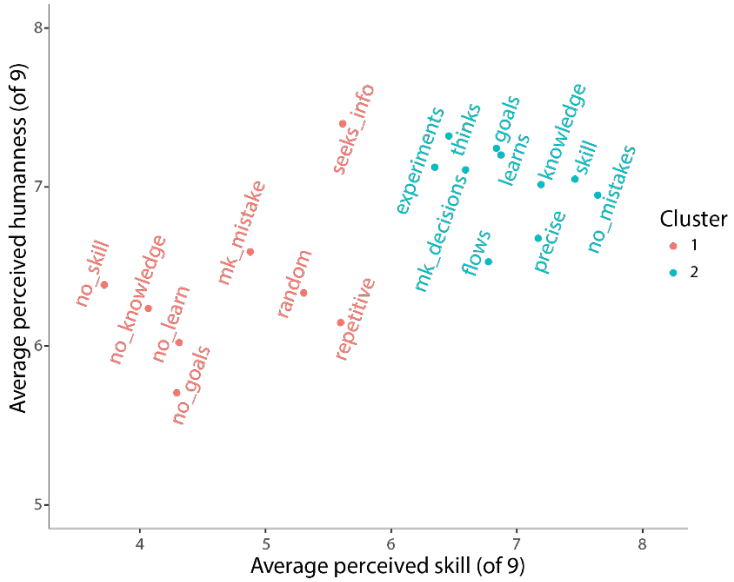

Figure 20: Character Traits by Average Perceived Human-Likeness and Average Perceived Predictability.

Statistics and Cluster Analysis for Game Theoretic Social Context

A X^2 analysis comparing the relationship between player and strategy was significant, $X^2 = (15,444) = 27.95$, $p = 0.022$. Figure 21 shows the proportion of participants who selected each GT game for each player.

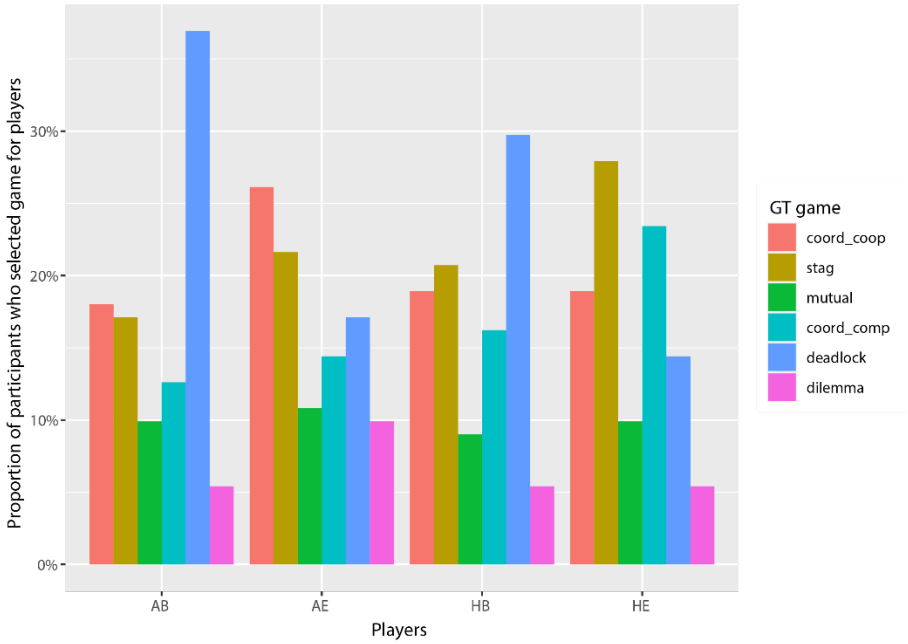

Figure 21: Percent of Participants who Selected Each GT Game for Each Player. The first three games (coord_coop, stag, and mutual) are cooperative games with different contexts and the last three games (coord_comp, deadlock, and dilemma) are competitive games with different contexts.

The ANOVA for average perceived human-likeness for GT game was not significant, $F(5,438) = 0.95$, $p = 0.448$, $\eta_p^2 = .011$, nor was the ANOVA for average perceived predictability for GT game, $F(5,438) = 2.22$, $p = 0.051$, $\eta_p^2 = 0.025$, though it was nearly significant. The ANOVA for average perceived skill was significant for GT game, $F(5,438) = 7.48$, $p < 0.001$, $\eta_p^2 = 0.079$. A post hoc Tukey's HSD showed that there were significant differences between deadlock (mean = 4.49, SD =2.87) and cooperative coordination (mean = 5.76, SD = 2.54; $p < 0.001$), deadlock and mutual assured destruction (mean = 5.69, SD = 2.31; $p = 0.003$), deadlock and stag hunt (mean = 5.66, SD = 2.43; $p = 0.001$), and cooperative coordination and social dilemma (mean = 4.62, SD = 2.92, $p = 0.041$). No other differences were statistically significant. Results can be seen in Figure 22.

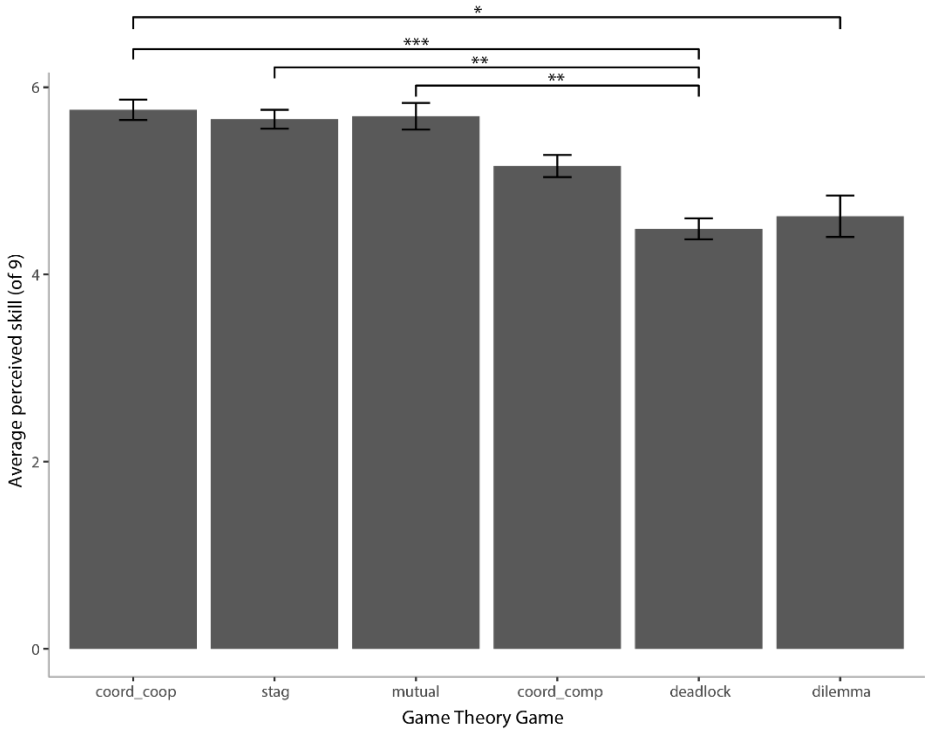

Figure 22: Average Perceived Skill by GT Game. Standard error bars are presented along with average perceived skill. The first three games (coord_coop, stag, and mutual) are cooperative games with different contexts and the last three games (coord_comp, deadlock, and dilemma) are competitive games with different contexts.

A cluster analysis was conducted for the GT game with respect to the averages of perceived human-likeness, skill and predictability. To do this, for every GT game, every block in which that GT game was selected was marked, and the average perceived human-likeness, skill and predictability for that block was included in an overall average of the three measures for that GT game.

First, the Euclidean equation was used to calculate the correlation distance between measures (see Figure 23).

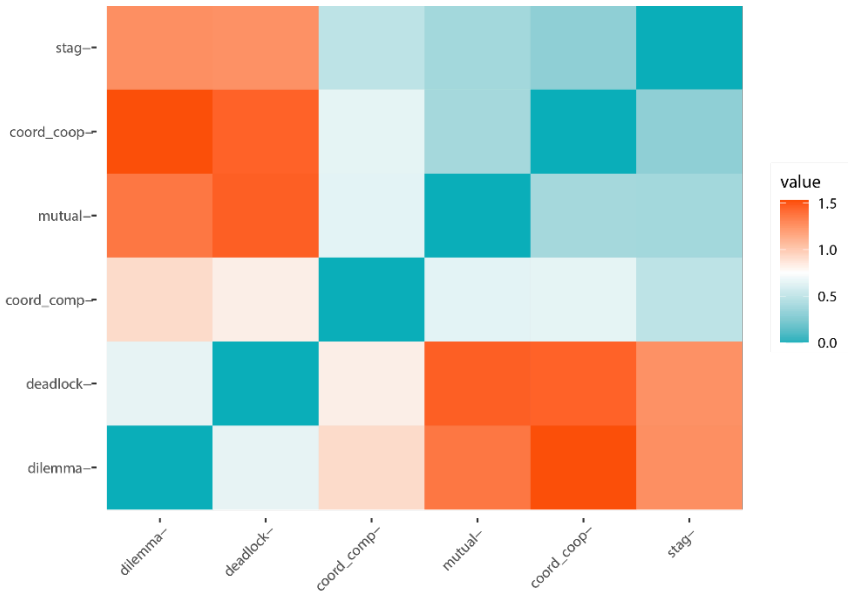

Figure 23: Results from Euclidian Distance Equation for GT Games. Higher values (in red) indicate larger distance between terms.

Next, the Elbow method was used to estimate the optimal number of clusters, and 2 was determined to be the best number (see Figure 24).

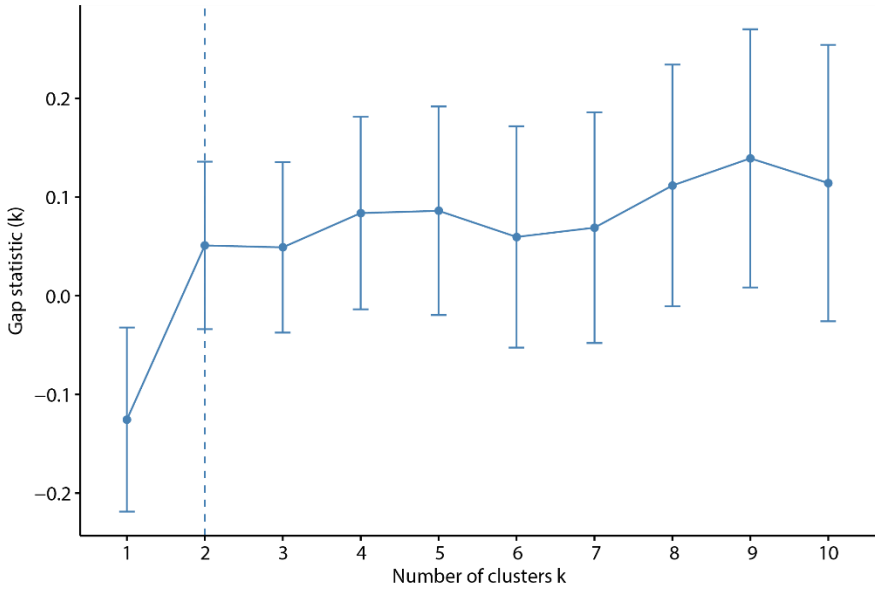

Figure 24: Optimal Number of Clusters According to Elbow Analysis for GT Games.

Next, a k-means clustering unsupervised learning algorithm was used to separate the GT games into groups (see Figure 25). The average perceived human-likeness is 6.68 for cluster 1 and 6.78 for cluster 2. Average perceived skill is 5.55 for cluster 1 and 6.57 for cluster 2. Average perceived predictability is 5.87 for cluster 1 and 6.51 for cluster 2. ANOVAs demonstrated that average perceived human-likeness by cluster was not significant, $F(1,4) = 0.19$, $p = 0.687$, part eta = 0.045, but average perceived skill by cluster was significant, $F(1,4) = 23.12$, $p = 0.009$, part eta = 0.852, and average perceived predictability by cluster was significant, $F(1,4) = 27.29$, $p = 0.006$, part eta = 0.872.

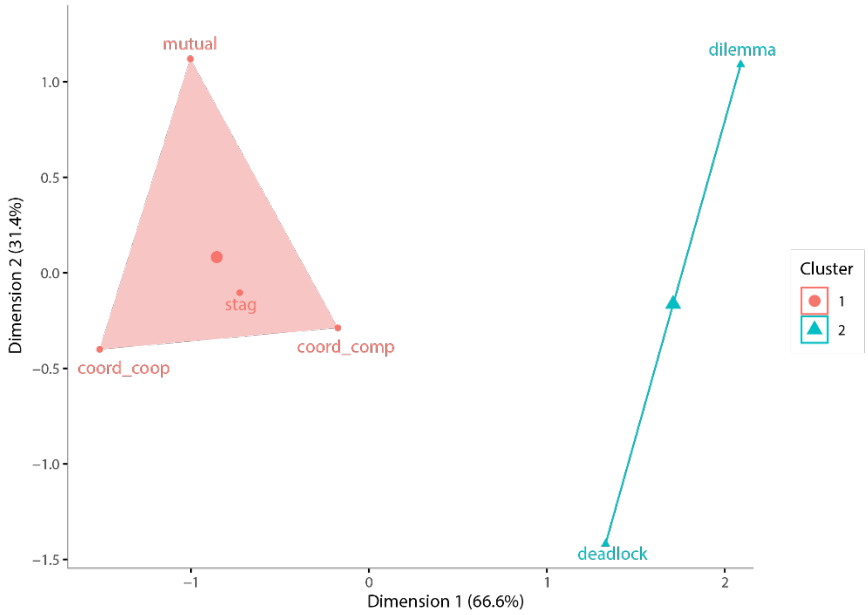

Figure 25: Cluster Plot of GT Games.

Finally, an individual plot was created for GT games along the dimensions of average perceived skill and average perceived predictability (Figure 26).

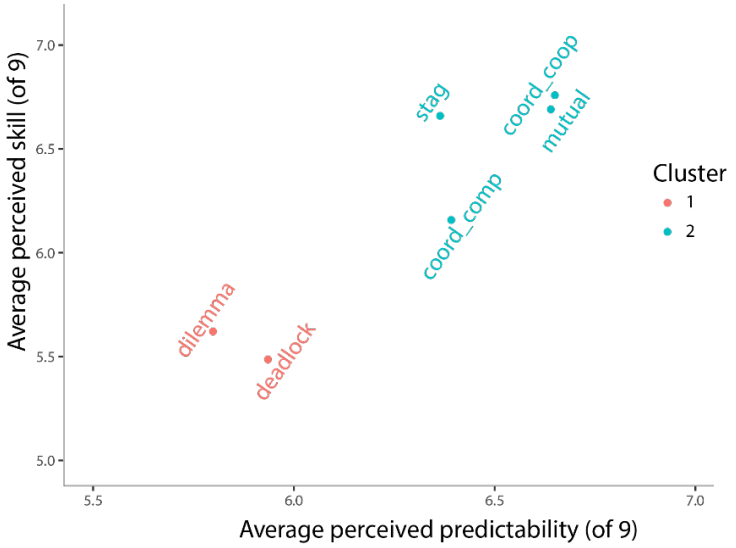

Figure 26: GT Games by Average Perceived Skill and Average Perceived Predictability.

Discussion

While H1 was not entirely supported, the relationship between competence (i.e., skill) and perceived human-likeness was non-linear. The trend indicated that overall, as perceived competence increased, so did perceived human-likeness, with the level of perceived human-likeness trending lower at the level of intermediate and perception of human-likeness was maximum at the highest level of competence (i.e., the opposite direction of what was hypothesized).

The observed relationship between average perceived human-likeness and average perceived skill may have been influenced by a couple factors. First, there was a high degree of variability in average perceived human-likeness below the level of expert. A visual inspection of the scatterplot of average perceived human-likeness by average perceived skill seemed to indicate that this variability may have been driven by perceived human-likeness. For instance, at a low level of perceived human-likeness, players had lower levels

of perceived skill and none were rated at the level of expert, whereas at a high level of perceived human-likeness, players' average ratings of perceived skill ranged from very low to very high. This pattern led to the post hoc analyses of perceived predictability by perceived skill with respect to 3 levels of human-likeness, which gave some indication that the relationship between perceived predictability and skill differed at the level of medium perceived human-likeness. While an obvious reason for this may be that it is harder to perceive behaviors of unknown origin as explainable/predictable, it is possible that some deeper reason is at play. For instance, it is possible that bottom-up perceptions of the three measures (human-likeness, skill and predictability) interact before a final perception is reached. Perhaps at a medium level of perceived humanness, there is no strong understanding of how to interpret motivations that lead to behavior (i.e., applying ToM), so the perception of an agent that jumps up and down is perceived as "random" rather than "having a goal in mind" (e.g., learning the controls or experimenting with the level). It is possible that if ToM was applied, the perception that the player had a goal could have influenced the final interpretation of human-likeness and skill. Future research would be needed to understand this relationship.

Second, there was also a bias towards higher ratings of skill and human-likeness. This was not entirely unanticipated, as stimuli was designed to be ambiguous. Ambiguous stimuli may have been more likely to be perceived as intermediate competence and middle level human-likeness. Additionally, vignettes were very brief (ranging from 3 to 8 seconds each), which may have also contributed to ambiguity.

H2 was also only partially supported. Beginner, intermediate and expert levels of competence did have an effect on the best fit trendlines for perceived human-likeness by perceived predictability, and the low and high end of predictability were associated with higher levels of perceived human-likeness for intermediate and expert levels, however, the beginner and expert levels did not show that pattern predicted. Additionally, there was a positive linear trend where increased predictability was related to increased perceived human-likeness, which was opposite of what was predicted.

Additionally, the trendline associated with the expert level had the maximum curvature that was expected of the intermediate level. However, H1 was that maximum perceived human-likeness should be associated with the intermediate level, and it was predicted that perceived human-likeness would decrease at the level of expert, but this was not the case. The maximum curvature associated with the level of expert is possibly due to this relationship. If this is the case, just like with H1, the fact that vignettes were very brief and intentionally ambiguous may have affected this relationship.

The cluster analysis of character traits showed that ToM traits seemed to have closer distance than traits that were not related to ToM states. ToM states were mostly clustered together (cluster 2), except for seeks information. The cluster that contained the ToM states also included the traits makes no mistakes, flowing, precise, and skill, and this cluster overall had significantly higher levels of perceived human-likeness, skill, and predictability. This is in line with the explanation that agents who were perceived as sufficiently human-like and sufficiently skilled demonstrated actions that were predictable

in the sense that participants could intuitively understand the motivations of the player by using ToM.

Overall, while some qualities match the category system from Experiment 1, others have shifted levels of perceived predictability or human-likeness. This may also relate to the fact that increased perceived predictability was associated with increased perceived human-likeness (which was again opposite from H2). One thing that stood out was the traits "making mistakes" and "making no mistakes". While in Experiment 1, "making mistakes" was often associated with human-likeness, and some participants expressed that AI should "make no or very few mistakes" (i.e., a bias in line with the Perfect Automation Schema), these patterns were not present in the cluster analysis. However, when looking at the character traits participants explicitly associated with humans or AI (i.e., results in Figure 15), these patterns were present as "makes mistakes" was more often associated with humans than AI and "making no mistakes" was more often associated with AI than humans. It is possible that this bias was present and affected some other perceptions of these players (for instance trust, which was not measured in this experiment), but that it did not actually affect perception of human-likeness during the experiment.

The analyses for the GT games demonstrated that player type did affect the types of game selected, as did average perceived skill, and average perceived predictability was almost significant, whereas average perceived human-likeness was not. Overall, average perceived human-likeness was not different across all agents, meaning that participants could not distinguish between AI and human players overall. It is interesting that even

while they were not able to reliably distinguish between these players, some of the GT games selected for these players had the same pattern as in Experiment 1. What is more interesting is that this pattern exits even though participants did not know the true identity of players (i.e., true human-likeness or true skill) when they selected games in this experiment but were able to make game selections for any hypothetical agent in Experiment 1 (see Table 3 and Figure 21). For instance, in this experiment and Experiment 1, deadlock was more frequently selected for beginner players, and was highest for the AB. Similarly, in both Experiments, Stag Hunt was the most frequently selected game for HE, and the highest frequency of Stag Hunt was for the HE. The cluster analysis revealed 2 clusters, where both coordination games (cooperative and competitive), Stag Hunt and Mutual Assured Destruction were clustered together, and the social dilemma and Deadlock were in a separate cluster. Clusters were distinctly split across perceived predictability and perceived skill.

Overall, results from Experiment 2 suggest that the answer to research question 2 (i.e., to what extent do competence and predictability contribute to the perception of human-like behavior) is that in general, increased perceptions of competence lead to an increased perception of human-likeness, but the relationship is non-linear, and the causal relationship is not fully understood (i.e., does perception of human-likeness increase perceived skill, is it the other way around, or does it depend on other factors?). Overall, a higher value of perceived predictability was associated with higher values of perceived human-likeness, but this pattern was different for different perceived levels of

skill/competence, such that at a higher level of perceived competence, high and low values of perceived predictability were associated with higher values of perceived human-likeness. This type of pattern was predicted (thought it was believed that perceptions of human-likeness would be maximized at the intermediate level). Additionally, from the cluster analysis, ToM states in general had higher levels of perceived human-likeness, skill, and predictability, suggesting that the actions of players who were perceived as sufficiently humanlike and sufficiently skilled were perceived as predictable because a participant could intuitively understand that player's motivations by using ToM. While that was the expectation prior to conducting this experiment, future research would be needed to specifically prove that theory.

EXPERIMENT 3: DON'T STARVE TOGETHER

What characteristics of social interactions do humans use to differentiate between humans and AI, and how do Human-Human Interactions and Human-AI Interactions (HAI) differ within a complex environment?

The purpose of the third experiment was to expand upon the findings from the pilot study (i.e., Experiment 1) to understand how people make distinctions between humans and AI when they are given the ability to interact however they want within a complex game environment. In particular, *what characteristics of social interactions do humans use to differentiate between humans and AI, and how do Human-Human Interactions and Human-AI Interactions (HAI) differ within a complex environment.* Participants played an open world survival style videogame with one of 4 possible types of humans and AI co-players.

In Experiments 1 and 2, the context of judgement was a typical psychological experiment, where participants watched vignettes and then responded about their perceptions of the agents. In the real world, judgements of others occur on the basis of interaction as well as observing behaviors in complex and dynamic contexts. According to the results presented in Pfeiffer et al., (2011), the context of an interaction can affect the way behavior is perceived, and in particular, whether or not predictable behavior is perceived as human-like. In this experiment, participants have the opportunity to observe a co-player's behavior and engage in social interactions with them.

Participants

In total, 114 undergraduate students (mean age: 20.7, SD: 3.69; 64 females) participated in this study. Overall, 26 participants were removed due to some glitches associated with the modifications made to the game or other technical difficulties (e.g., game crashing after the experiment was started, internet connectivity issues, agent continuing to run into a wall or standing still for more than half the experiment), or incomplete datasets leaving early, and neglecting to fill out some questionnaires, which resulted in a total 88 usable data sets (mean age: 20.8, SD: 4.19; 47 female). The only screening criteria were that participants had no prior experience playing the game and were over the age of 18. Participants reported spending an average of 5.6 hours (SD = 2.70) on a computer per day and an average of 5.28 hours (SD = 8.16) playing videogames per week.

The number of participants collected for this study is consistent with, if not greater than, similar studies involving actual robots or AI (Salem et al., 2013; Wykowska et al., 2015; Hayes et al., 2014; Mutlu et al., 2009, etc.). Smaller sample sizes are generally accepted due to the difficulty of experimental set up and the frequent removal of participants due to technical difficulties. A post hoc power analysis was conducted to determine the power associated with the number of participants included in this study. With a medium effect size f2(V) = 0.154, α error probability of 0.05, n = 88, groups = 4, and 24 response variables, the observed power = 0.76. An attempt was made to collect more data, but George Mason University shut down due to COVID-19 before any additional data could be collected.

Stimuli

Four different agents were selected for the Turing Test, who participants played with in the game Don't Starve Together: The two human co-player types included a Confederate Human co-player who was given no instructions other than to play the game however he wanted to. The second human co-player type was another participant, such that two participants unwittingly played in tandem and judged the other participant. The two AI co-players included a Simple AI and a Social AI. The Simple AI was programmed to display human-like actions during game play (collecting resources, fighting or running away from monsters, interacting with a participant when they engaged with it), but had no incentives during the game. The behavior tree for the Simple AI co-player can be seen in Figure 27. The Social AI was developed with a cognitively plausible behavior tree that allowed it to use human-like motivations for survival and interaction behaviors within the game, as well as a machine learning algorithm to learn from social interactions and decide for itself how to treat participants in the game. This agent played the game more competently than the Simple AI in the sense that it actually tried to survive and interact, rather than just imitating common human behavior in the game. The behavior tree for the Social AI co-player can be seen in Figure 28.

The two avatars that players and co-players used were "Wilson" and "Woodie"; see Figure 29. These two avatars were selected for their similar features and relatively normal appearances compared to other available avatars in the game. In the AI conditions, the participant played as Wilson and the AI co-player played as Woodie. The avatar assignment was specified to make the start up procedure as straightforward as possible for

research assistants, as improperly starting the game with the re-search modifications or improperly adding the co-player to the game could cause the game to crash or result in missing data.

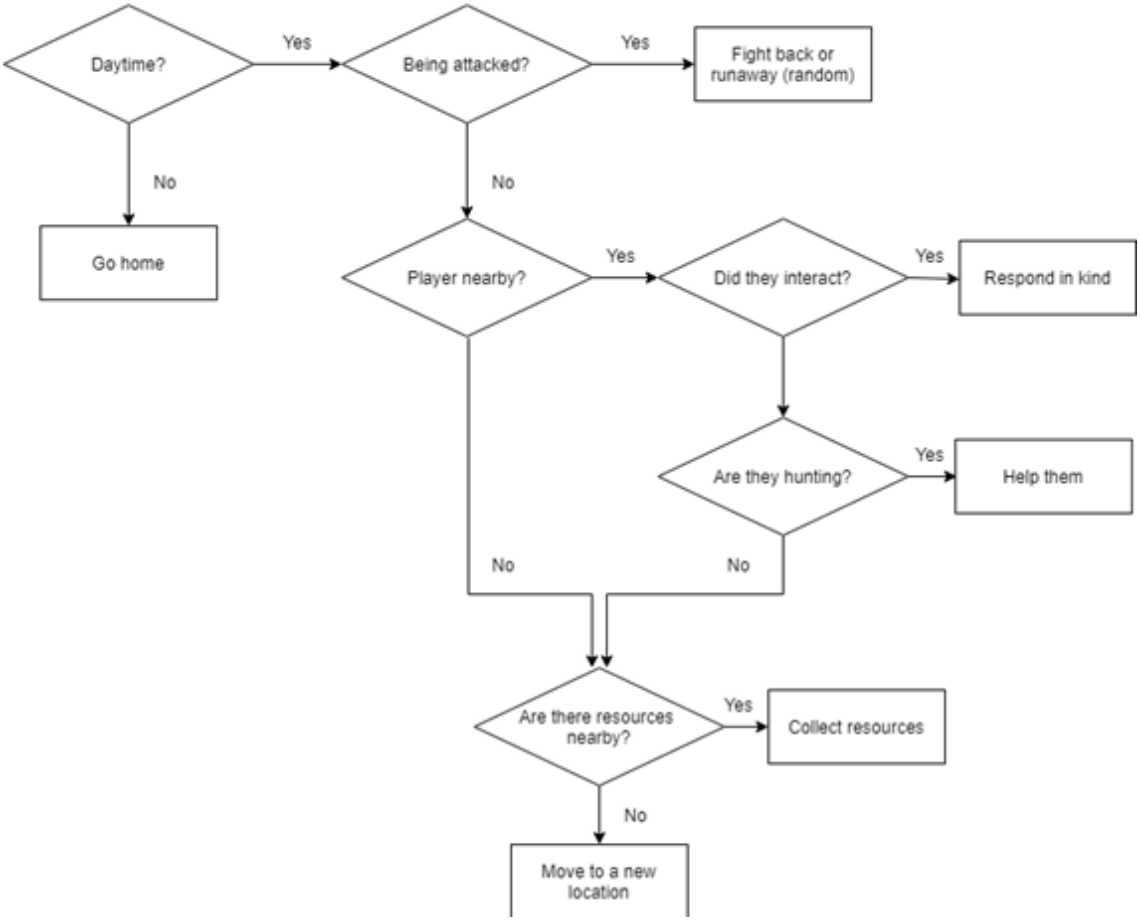

Figure 27: The Simple AI's Behavior Tree. The Simple AI's behavior was governed by this behavior tree.

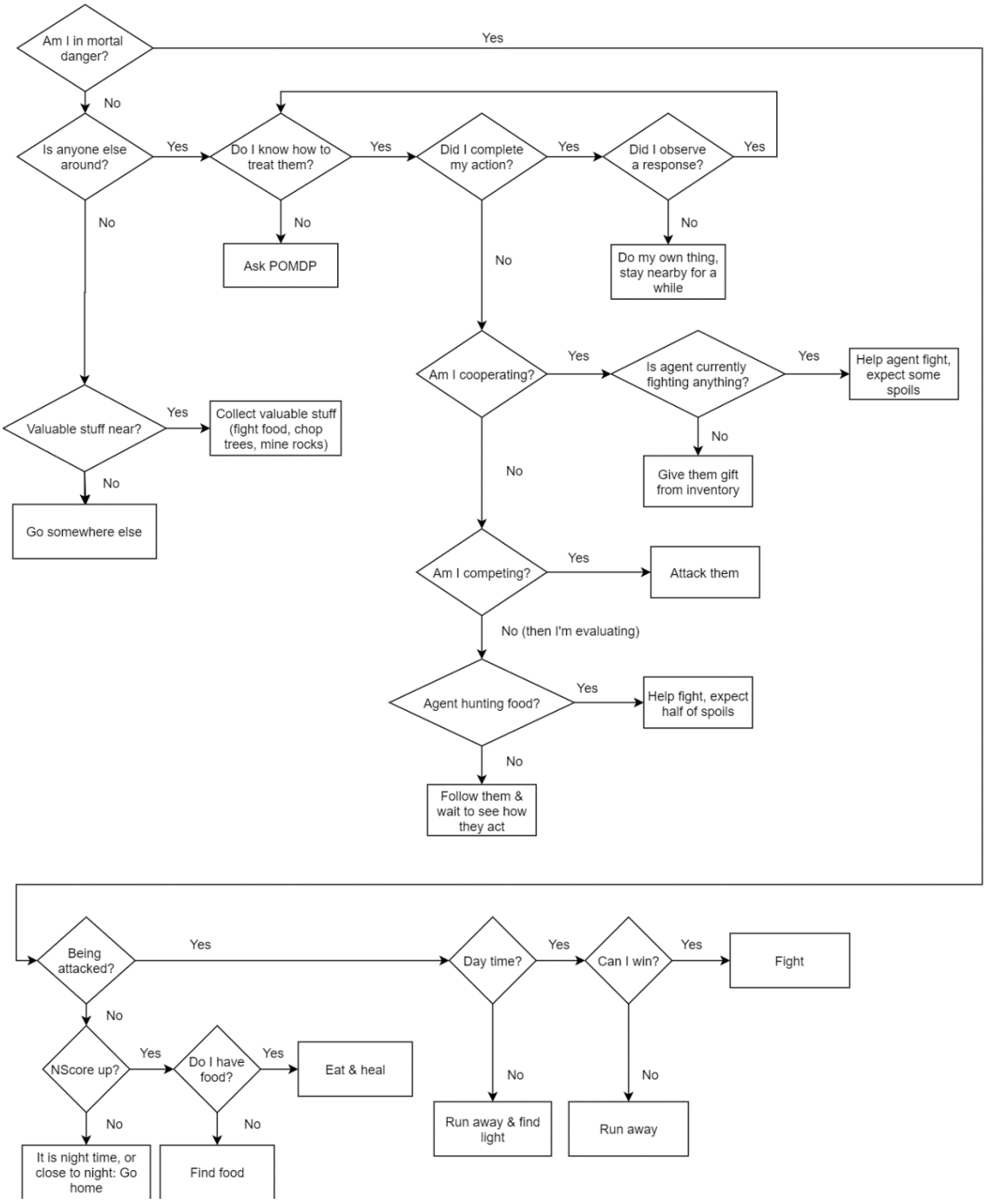

Figure 28: The Social AI's Behavior Tree. The Social AI's behavior consisted of two main components. The first component was a behavior tree that was designed from participant's statements about what constituted humanlike behavior in the game. The Social AI kept track of its own "neediness" based on its current player stats (health, hunger and sanity) and how many resources it had in inventory, or how well it was currently surviving in the game and made decisions based on how needy it was in the moment. The second component involved a POMDP (Emami, Hamlet &

Crane, 2015) that kept a memory of other agents it interacted with, estimated for itself the social context, and made decisions based on how it was being treated by the participant.

Figure 29: DST Player Avatars: Wilson (left) was played by the participant; Woodie (right) was played by half of the participants in the human condition and by the Confederate Human, Simple AI and Social AI.

Apparatus

Two copies of the game Don't Starve Together were purchased and modified to record participants' in-game behavior and interactions with co-players. Modifications were also made to make the game a little easier for the participant (e.g., they could never actually die, but were not informed of this fact), and the chat function was disabled to ensure that all interactions were behavioral. The game was played on PCs through the Steam gaming platform (Valve, 2003). Participants were given the option to use either an Xbox style controller or mouse and keyboard. All questionnaires were administered through Google Forms. Interviews were conducted verbally and transcribed by the researcher.

Measures

While playing the game, various behavioral measures are recorded associated with the participants' in-game behaviors, performance, and interactions with co-players: 1) distance between player avatars within the game environment (measured in approximate centimeters on the monitors), 2) how often participants gave / received items to / from co-players, and 3) how often participants attacked co-players / were attacked by co-players.

Participants also complete four surveys during the study. Prior to playing the game, participants filled out a generic demographics survey. After playing the game, participants filled out a series of questions to indicate how much they enjoyed their interactions in the game and with the co-player, how much they liked and trusted the co-player, and how much they would like to play an-other game in the future with the co-player , how much they would prefer to play cooperatively and how much they would prefer to play competitively with the co-player, a presence questionnaire to indicate what extent did they feel like the game and the other player were "real" (Schneider et al., 2004), and the Godspeed questionnaires of anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety (Bartneck et al., 2009) to indicate how participants perceived each co-player after their interactions.

After playing the game and completing the questionnaires, participants were interviewed with the following questions:

1. What agent do you think you were playing with?

2. On a scale from 1-10, with 1 being the least, 10 being the most, how confident are you that the agent was (Human/AI)?
3. What experience level do you think the agent has (beginner, intermediate, expert) Why?
4. On a scale from 1-10, with 1 being the least, 10 being the most, how "socially intelligent" (i.e., able to understand your intentions, knowingly act in a social way, or able to decide whether or not to cooperate with you) do you think the other agent was?
5. What made you think the agent was or was not "socially intelligent"?
6. What features made you think it was a (human/AI) player? Why?
7. What features of the performance made you think it was NOT a (human/AI)? Why?
8. What do you think the agent's goal was while they were playing?
9. If you were going to play another game with this agent, would you prefer a cooperative or competitive game, and why?

While the interview questions were preplanned, research assistants would often ask follow-up questions to encourage participants to elaborate on their responses to collect as much qualitative data as possible.

Procedure

At the beginning of the experiment, participants read the consent form and verbally confirmed that they consented to be in the study, then filled out the demographic questionnaire and were given instructions for the experiment. In the experiment, participants play Don't Starve Together, an immersive, multiplayer wilderness survival game in which players collect resources from the environment (e.g., food and fire wood) and craft tools and other objects (e.g., hats, armor, hand tools) to survive. Figure 30 shows a typical game view. The game operates on a day cycle, with a clock at the top right side of the screen indicating when it is morning, evening or night. At night, the entire field turns dark, and players must find light to see their environment and stay alive. Players can track how well they are performing by looking at their health, hunger, and sanity levels that are displayed on three icons in the top right of the screen, just under the clock. Players can also see how many items they have stored in their inventory (displayed at the bottom of the screen) and can interact with the crafting tab displayed on the left side of the screen to determine what items they can build given the resources carried in their inventory.

Figure 30: A Typical DST Game Scene: Both players (Wilson and Woodie) are in view. Both players are within the "home base" that featured some barriers, a cook pot and a constant light source that offered protection at night. The game statistics (hunger, health, sanity) are displayed in the upper right corner of the screen. The player's inventory is displayed on the bottom of the screen. On the left side of the screen is the "crafting tab" where players can view and pick from different recipes to build different items to aid in survival in the game once they have collected the necessary resources.

Players can play the game however they desire, including exploring the vast environment, fighting or befriending creatures found in the environment, or building equipment to help them survive and progress in the game.

Participants were then told that they would be playing Don't Starve Together with another player, and were instructed that they could do whatever they wanted in the game and towards the co-player (including exchanging goods like food, clothing, tools; fighting one another, and assisting each other in hunting food or fighting monsters), and that at the end we would ask them if they believed that the co-player was a human or an AI agent. They were also informed that chat within the game was disabled, and that all communication within the game would be behavioral.

After a 5-minute practice play, where participants are given some tips on how play the game and were allowed to ask questions, participants were then asked to leave the computer area while the experimenter brought the co-player into the game by initializing their avatar in the home base. In the confederate and tandem conditions, an online server was created on one lab computer where Wilson was selected as the player avatar, and the researchers communicated via text messaging on their phones to connect a secondary computer to the server and initialize Woodie, who was played either by the confederate co-player or another participate in the tandem condition. In both AI conditions, an online server was created with Wilson as the participant's avatar, and a new instance of the AI co-player (Simple AI or Social AI) was created with Woodie as the avatar. In the Social AI condition, the POMDP was running in a terminal in the background, and was not visible to the participant at any point. Once both players were initialized within the environment, the participant played the game with their co-player for approximately 30 minutes. During the experiment, data was saved after every interaction participants have with the environment and with the co-player for subsequent analysis. After playing for 30 minutes, the game was turned off and participants filled out the Godspeed measure, the presence questionnaire and the enjoyment questionnaire. Finally, participants were verbally interviewed about the perceived identity of the co-player and their overall impressions. At the end of the experiment, participants were told the true identity of their co-player and thanked for their participation. The experiment took about 1 hour to complete.

Quantitative Analysis

In-game distance between players (on-screen distance between player avatars) and in-game behaviors (how often players interacted with each other by exchanging gifts, helping each other in combat or hunting, or attacking each other) were compared across co-player identity (i.e., Human-likeness) as well as perceived identity (i.e., Perceived Human-likeness). All survey measures were also compared across Human-likeness and Perceived Human-likeness.

As in Experiment 1, performance on the Turing Test was evaluated by comparing the relative frequency that co-players were rated as humans. Accuracy in detecting Human-likeness was compared against chance (similar to Wykowska et al., 2015) to estimate how sensitive participants were to human-likeness in the experiment.

Data was analyzed using a MANOVA with co-player identity and perceived human-likeness of co-player as IVs and the 6 behavioral measures (average in-game distance between players, health, hunger, sanity, and total times the participant interacted with, and was interacted with by the co-player) and 16 questionnaire measures (to what extent did participants feel like the game and co-player were "real", how much participants liked and trusted co-players, how much they enjoyed the interaction, would like to play again in the future with the co-player, how much they would like to play cooperatively and competitively with the co-player, averages on each of the Godspeed questionnaires: anthropomorphism, animacy, likability, perceived intelligence and perceived safety) as DVs (22 total DVs).

Qualitative Analysis

Participants were interviewed about their perceptions of the co-player and their natural language responses were analyzed and coded by two raters to determine what cues participants used to determine human-likeness in the game. The same bottom-up procedure was used to generate the category system in this experiment as was used in Experiment 1. Given the amount of quantitative data in this experiment, only question 6 ("What features made you think it was a (human/AI) player?") was analyzed in the qualitative analysis. A consensus between the two raters was obtained using the same procedure as in Experiment 1 and inter-rater reliability was calculated in the same way with Cohen's Kappa and χ^2 between the two raters. Using the same procedure as in Experiment 1, an empirical game theoretic analysis was performed on participants' responses to how they would prefer to play with the same co-player in the future (cooperatively or competitively) and their justifications for that decision.

Results

Qualitative Results

Overall, the Confederate Human co-player was judged as a human by 18.2%, SD = 38.6%, of participants, the Tandem Human co-player as a human by 45.5%, SD = 49.8%, of participants, the Simple AI co-player a human by 22.7%, 41.9%, of participants, and the Social AI co-player as a human by 31.8%, SD = 46.6%, of participants. Overall, 70% of co-players were rated AI and 30% were rated human. Every co-player was rated as an AI more than 50% of the time. Accuracy in detecting humanness for the Confederate Human

was significantly below chance (18.2% accurate, t(21) = -3.78, p < .001). Accuracy in detecting humanness for the Tandem Humans was not significantly below chance (45.5% accurate, t(21) = -0.42, p = 0.340). Accuracy in detecting humanness for the Simple AI was significantly above chance (77.3% accurate, t(21) = 2.98, p = 0.004), as was accuracy for the Social AI (68.2% accurate, t(21) = 1.79, p =.044), all one-tailed.

The Confederate Human co-player was explicitly rated as an expert more often than any other player (36%, SD = 48.1%, of participants), and beginner least often of any other player (9%, SD = 28.7% of participants), while the Simple AI was rated as a beginner most often (55%, SD = 49.8%, of participants), and the Tandem Human co-players and Social AI co-players were rated as intermediates most often (68%, SD = 46.6%, of participants and 55%, SD = 49.8%, of participants, respectively). Figure 6 shows perceived human-likeness by calculated expertise with a second order polynomial best fit line. As seen in Figure 31, perceived human-likeness seems to peak at around 40%.

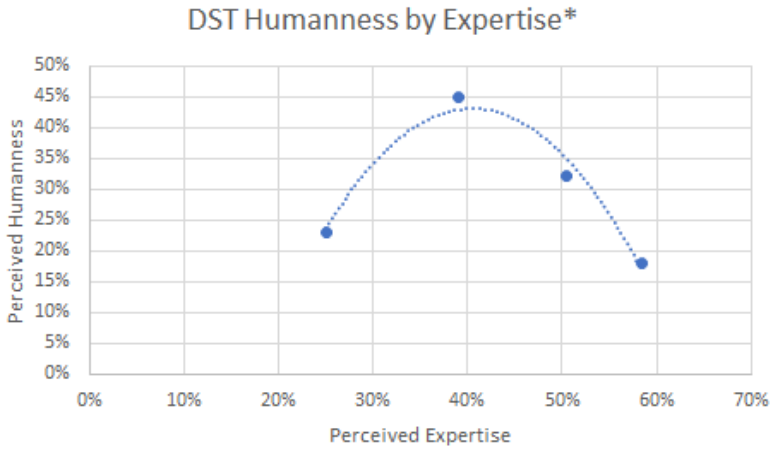

Figure 31: Explicit Ratings of Human-Likeness by Expertise in The DST Experiment. Similar to Figure 6 and the MarI/O experiment, a second order polynomial best fit line has been added and shows that perceived human-likeness

peaks near mid-level expertise. *Since participants were allowed to rate the co-player as a beginner, intermediate or expert, expertise is calculated by (%_rated_beginner * 0 + %_rated_intermediate * 0.5 + %_rated_expert * 1).

MANOVAs were conducted for behavioral and questionnaire data. A MANOVA was conducted with co-player identity and perceived human-likeness as IVs and 22 behavioral and questionnaire measures as DVs. The multivariate result was significant for co-player identity, Pillai's Trace = 1.80, $F = 4.04$, $df = (3,78)$, $p < 0.001$, indicating that there were differences in participant perceptions and interactions with different co-players. The significant univariate effects for co-player identity are presented in Table 6. While a number of measures varied significantly by co-player identity, of note are the number of times players and co-players interacted with one another, which is highest for the Social AI, the ratings of how much participants trusted co-players (Figure 32a), liked co-players (Figure 32b), and perceived the co-player as a "real person" (Figure 33a). There were also significant differences in mean ratings on four of the five Godspeed measures (Animacy, Intelligence, Likability, and Anthropomorphism).

A post hoc Tukey's HSD showed that there were significant differences in ratings of how much co-players were trusted between the Confederate Human (mean = 5.23, SD = 2.31) and the Social AI (mean = 2.36, SD = 1.47; $p < 0.001$). No other differences were significant, though the difference between the Confederate Human and Simple AI (mean = 3.45, SD = 2.61) was nearly significant ($p = 0.059$), as was the difference between the Tandem Human (mean = 4.14, SD = 2.62) and the Social AI ($p = 0.059$). Results can be seen in Figure 32a.

A post hoc Tukey's HSD showed that there were significant differences in ratings of how much co-players were liked between the Confederate Human (mean = 5.45, SD = 1.99) and the Social AI (mean = 3.55, SD = 1.57; p = 0.011). No other differences were significant. Results can be seen in Figure 32b.

A post hoc Tukey's HSD showed that there were significant differences in ratings of co-players as a "real person" between the Simple AI (mean = 3.27, SD = 2.25) and the Social AI (mean = 5.14, SD = 2.29; p = 0.032). No other differences were significant, though the difference between the Tandem Human (mean = 5.00, SD = 2.25) and Simple AI was nearly significant (p = 0.054). Results are shown in Figure 33a.

The MANOVA result was also significant for perceived human-likeness, Pillai's Trace = 0.69, F = 5.90, df = (1,78), $p < 0.001$, indicating that participants' perceptions of co-players and the interaction was impacted by whether or not participants believed the co-player was a human. The significant univariate effects for perceived human-likeness are presented in Table 7. Mean ratings of how much the co-player was perceived as a "real person" is shown in Figure 33b. There were also significant differences in mean ratings on two of the five Godspeed measures (Anthropomorphism and Animacy).

Table 6: Significant Univariate Effects for Co-player Identity.

Dependent Variable	df	df error	F	p		Co-player	Means	95% Confidence Interval Lower bound	95% Confidence Interval Upper bound
Average times co-player interacted	3	82	26.86	< 0.001	***	Confederate	0.62	-4.91	6.14
						Tandem	0.36*	-5.04	5.77
						Simple AI	0.59	-4.81	6.15
						Social AI	**29.00**	23.59	34.40
Average times participant interacted*	3	82	15.96	< 0.001	***	Confederate	1.86	-3.47	7.18
						Tandem	1.09*	-4.11	6.29
						Simple AI	1.45	-3.75	6.65
						Social AI	**22.41**	17.21	27.61
Average distance between players **(approx. cm.)**	3	82	14.51	< 0.001	***	Confederate	**154.38**	129.73	179.03
						Tandem	109.48	85.40	133.56
						Simple AI	40.93	16.85	65.01
						Social AI	86.98	62.90	111.07
Average participant health	3	82	13.28	< 0.001	***	Confederate	**66%**	56%	77%
						Tandem	43%	33%	54%
						Simple AI	61%	51%	71%
						Social AI	24%	13%	34%
Perceive co-player as "real person"	3	82	7.53	< 0.001	***	Confederate	3.67/7	2.71	4.63
						Tandem	5.00/7	4.06	5.94
						Simple AI	3.27/7	2.34	4.21
						Social AI	**5.14**/7	4.20	6.07
Trust co-player	3	82	6.25	< 0.001	***	Confederate	**5.33**/7	4.33	6.33
						Tandem	4.14/7	3.16	5.11
						Simple AI	3.45/7	2.48	4.43
						Social AI	2.36/7	1.39	3.34
Mean GS Animacy	3	82	4.11	0.009	**	Confederate	2.78/5	2.42	3.13
						Tandem	2.89/5	2.55	3.24
						Simple AI	2.23/5	1.88	2.58
						Social AI	**2.90**/5	2.55	3.25
Confidence in Turing response	3	82	4.03	0.010	*	Confederate	6.38/10	5.66	7.10
						Tandem	6.33/10	5.62	7.05
						Simple AI	**7.68**/10	6.98	8.38
						Social AI	6.18/10	5.48	6.88

Mean GS Intelligence	3	82	3.88	0.012	*	Confederate	**3.70**/5	3.30	4.09
						Tandem	3.33/5	2.94	3.72
						Simple AI	2.81/5	2.42	3.20
						Social AI	2.95/5	2.57	3.34
Like co-player	3	82	3.84	0.013	*	Confederate	**5.48**/7	4.61	6.35
						Tandem	4.14/7	3.29	4.99
						Simple AI	4.45/7	3.61	5.30
						Social AI	3.55/7	2.70	4.39
Mean GS Likability	3	82	3.83	0.013	*	Confederate	**3.18**/5	2.87	3.49
						Tandem	2.87/5	2.57	3.17
						Simple AI	2.85/5	2.55	3.15
						Social AI	2.45/5	2.15	2.75
Mean GS Anthropomorphism	3	82	3.68	0.016	*	Confederate	2.70/5	2.35	3.04
						Tandem	2.78/5	2.44	3.12
						Simple AI	2.24/5	1.90	2.57
						Social AI	**2.85**/5	2.52	3.19
Felt like they were "really there"	3	82	3.21	0.027	*	Confederate	5.67/7	4.69	6.64
						Tandem	4.86/7	3.91	5.82
						Simple AI	5.00/7	4.05	5.95
						Social AI	**6.50**/7	5.55	7.45
Average participant sanity	3	82	3.06	0.033	*	Confederate	**83%**	75%	92%
						Tandem	67%	59%	75%
						Simple AI	82%	74%	91%
						Social AI	80%	72%	88%
Average participant hunger	3	82	2.92	0.039	*	Confederate	66%	60%	72%
						Tandem	56%	50%	62%
						Simple AI	64%	58%	71%
						Social AI	**69%**	63%	76%

* Interactions from participant included times in which participants "gave" the co-player items by placing them at the feet of the co-player (less than 7 approx. cm away on the screen). Many participants expressed that they did this instead of using the "give" function that was programmed specifically for this interaction. While this added to the number of times participants interacted, it is not considered in the number of co-player interactions in the tandem condition, as it is less salient than the "give" function and participants may not have noticed this.

Table 7: Significant Univariate Effects for Turing Test Response (i.e., Perceived Human-likeness)

Dependent Variable	df	df error	F	p		Response	Means	95% Confidence Interval Lower bound	95% Confidence Interval Upper bound
Perceive co-player as "real person"	1	84	70.79	< 0.001	***	Human	**6.80/7**	6.13	7.47
						AI	3.26/**7**	2.83	3.68
Mean GS Anthropomorphism	1	84	37.56	< 0.001	***	Human	**3.30/5**	3.01	3.58
						AI	2.38/**5**	2.20	2.56
Mean GS Animacy	1	84	21.45	< 0.001	***	Human	**3.30/5**	3.00	3.60
						AI	2.46/**5**	2.26	2.65
How much participants want to play again	1	84	10.74	0.002	**	Human	**5.00/7**	4.01	5.99
						AI	3.44/**7**	2.81	4.06
Average participant sanity	1	84	4.53	0.036	*	Human	69%	61%	77%
						AI	**82%**	77%	87%

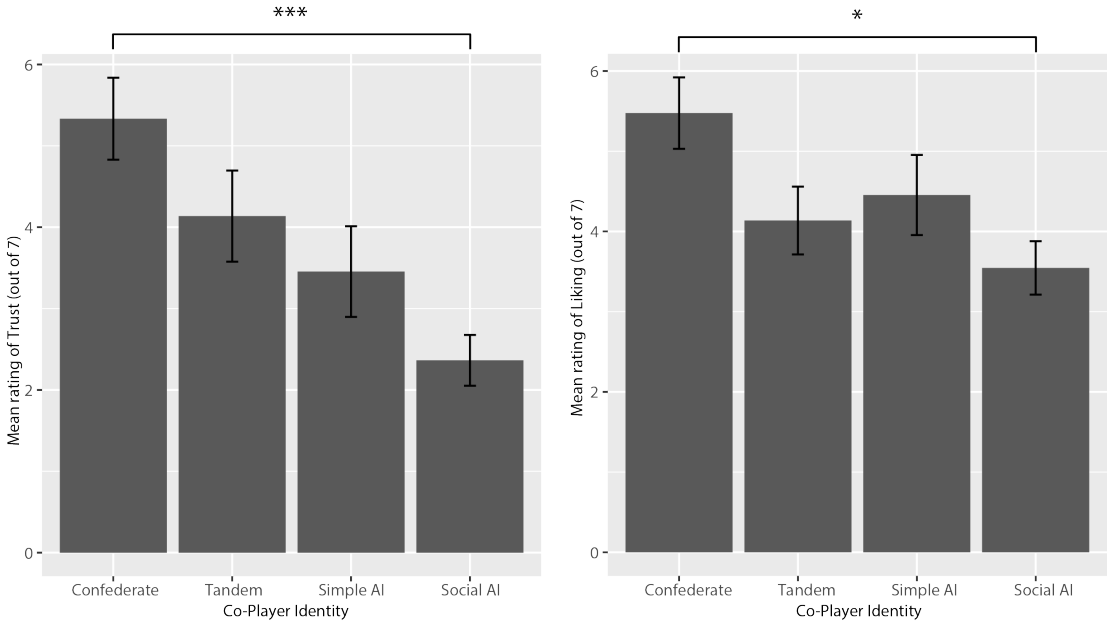

Figure 32: Mean Ratings of How Much Participants Trust (Left) and Like (right) the Co-Players. The Confederate Human co-player was trusted and liked more than any other co-player.

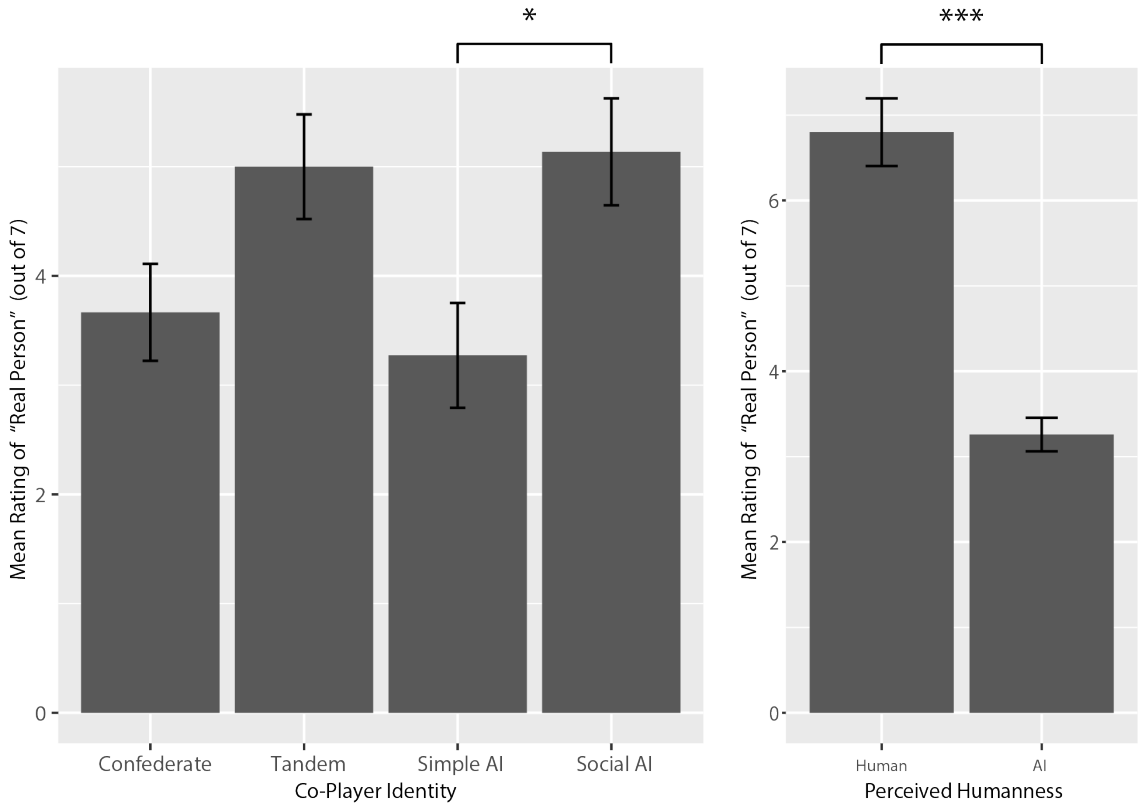

Figure 33. Mean Participant Ratings of Co-Players as a "Real Person" for Co-Player Identity (Left) and Perceived Identity (Right). The Social AI received the highest mean ratings of being perceived as a "real person" and the confederate Simple AI received the lowest mean ratings. Participants who perceived the co-player as a human also reported higher mean ratings of perceiving co-players as a "real person".

Quantitative Results

Results from the qualitative analysis are presented in Table 8. Eighty-eight statements were analyzed. The same qualitative analysis process was followed for Experiment 3 as was in Experiment 1. While some cues/categories are the same as in Experiment 1, many are unique to Experiment 3 because differences in paradigm (i.e., observation vs. interaction) led participants to use different cues. Since the amount of qualitative data analyzed was much smaller than Experiment 1, and the intent was to focus more on quantitative data, no further attempt was made to determine mid-level or overarching categories. As a follow up

analysis, χ^2 tests were used to provide a look at what cues led to the perception of human-likeness and were associated with which co-players. The table is sorted by overall cue (i.e., category) frequency, with the most common cue occurring 31 times and the least common cue occurring 2 times. χ^2 results were significant for movement, interaction between players, and acting cooperatively with respect to perceived human-likeness, where interaction between players and acting cooperatively were more associated with perceived human-likeness and mentioning qualities of movement was more associated with the perception of AI-ness. χ^2 results were significant for random or unpredictable behavior and too much interaction between players with respect to co-player identity, where both qualities were more associated with the Social AI.

The mid-level and overarching categories from Experiment 1 were applied, but the value in mid-level categories especially was limited as most cues were based on the direct observation of behavior without a lot of interpretation provided within question 6. For most cues, predictability and unpredictability here seemed to be more related to whether the co-player's behavior violated what was expected of human co-players. Results are shown in Table 9.

Table 8: Summary of Coded Statements from Qualitative Analysis. The table shows the cue participants used, a representative example, the total count of coded statements associated with the cue and relative frequency that it was mentioned by all participants, Cohen's Kappa as a measure of inter-rater reliability and a χ^2 between raters, and counts, relative frequencies, and χ^2 tests broken down by perceived human-likeness and co-player identity.

					count (and % of participants who mentioned cue)			count (and % of participants who mentioned cue)				
cue	**example**	**count (and % of participants who mentioned cue)**	**Cohen's Kappa**	**χ^2 between raters**	**perceived as human**	**perceived as AI**	**χ^2 for perceived humanness**	**Confederate human**	**Tandem human**	**Simple AI**	**Social AI**	**χ^2 for Co-player identity**
not enough interactions with participant	"When the game started, they immediately walked away. I would expect that a person would try and tell if I was hostile or friendly."	31 (35%)	0.87	(1,88) = 63.04, p <0.001	6 (23%)	25 (40%)	(1,88) = 1.42, p = 0.234	9 (41%)	7 (32%)	10 (45%)	5 (23%)	(1,88) = 2.84, p = 0.417
movement	"They were very active and their movement was fluid"	18 (20%)	0.678	(1,88) = 35.95, p <0.001	2 (8%)	20 (32%)	**(1,88) = 4.34, p = 0.037**	8 (36%)	3 (14%)	8 (36%)	3 (14%)	(1,88) = 5.78, p = 0.123
interacted with environment	"He knows what to do in this game. Everytime I saw him, he kept working."	15 (17%)	0.649	(1,88) = 31.33, p <0.001	6 (23%)	9 (15%)	(1,88) = 0.56, p = 0.456	4 (18%)	5 (23%)	4 (18%)	2 (9%)	(1,88) = 1.69, p = 0.64
skilled	"They were getting resources and building stuff, like they knew what to do in the game."	15 (17%)	0.767	(1,88) = 46.05, p <0.001	4 (15%)	11 (18%)	(1,88) = 0, p = 1	5 (23%)	5 (23%)	3 (14%)	2 (9%)	(1,88) = 2.32, p = 0.508
interacted with participant	"helped me chase the rabbit and helped with other tasks"	14 (16%)	0.767	(1,88) = 45.18, p <0.001	10 (38%)	5 (8%)	**(1,88) = 10.59, p = 0.001**	3 (14%)	5 (23%)	3 (14%)	4 (18%)	(1,88) = 1.05, p = 0.789
not enough interactions with environment	"It didn't collect the logs after the trees were cut down"	12 (14%)	0.629	(1,88) = 28.39, p <0.001	1 (4%)	11 (18%)	(1,88) = 1.79, p = 0.181	0 (0%)	3 (14%)	4 (18%)	5 (23%)	(1,88) = 5.36, p = 0.147
random/unpredictable behavior	"Player 2 would attack for no reason"	12 (14%)	0.705	(1,88) = 37.54, p <0.001	4 (15%)	8 (13%)	(1,88) = 0, p = 0.972	1 (5%)	2 (9%)	0 (0%)	9 (41%)	(1,88) = 19.03, p = 0
unskilled	"The way he was moving, there was obvious confusion. Seems he was trying to learn something"	8 (9%)	0.377	(1,88) = 10.62, p = 0.0011	3 (12%)	5 (8%)	(1,88) = 0.03, p = 0.869	1 (5%)	1 (5%)	5 (23%)	1 (5%)	**(1,88) = 6.46, p = 0.091**
too much interaction with participant	"just following me around and attacking as often as possible until I attacked him back once or twice."	8 (9%)	0.751	(1,88) = 39.05, p <0.001	3 (12%)	5 (8%)	**(1,88) = 0.03, p = 0.869**	0 (0%)	0 (0%)	2 (9%)	6 (27%)	(1,88) = 12.97, p = 0.005
cooperative behavior	"The fact he interact with me, he gave me so many things"	5 (6%)	0.739	(1,88) = 34.53, p <0.001	4 (15%)	1 (2%)	(1,88) = 4.41, p = 0.036	1 (5%)	2 (9%)	0 (0%)	2 (9%)	(1,88) = 2.41, p = 0.492
weird interactions with environment	"The continuous movement and goal satiation. He never stopped. He ran far out too."	3 (3%)	0.649	(1,88) = 24.93, p <0.001	0 (0%)	6 (10%)	(1,88) = 1.31, p = 0.252	3 (14%)	2 (9%)	1 (5%)	0 (0%)	(1,88) = 3.6, p = 0.308
learning	"followed around, mimicked actions, and was observant of me from time to time"	3 (3%)	-0.0156	(1,88) = 0, p = 1.000	2 (8%)	1 (2%)	(1,88) = 0.69, p = 0.408	1 (5%)	1 (5%)	1 (5%)	0 (0%)	(1,88) = 1.05, p = 0.788
repeatitive	"It was very predictable when he would come up to me"	2 (2%)	0.661	(1,88) = 10.25, p <0.001	0 (0%)	2 (3%)	(1,88) = 0.01, p = 0.906	0 (0%)	0 (0%)	1 (5%)	1 (5%)	(1,88) = 2, p = 0.572
competitive behavior	"the other agent played more independently after I ignored him. He also took my berries."	2 (2%)	0.661	(1,88) = 10.25, p <0.001	1 (4%)	1 (2%)	(1,88) = 0, p = 1	0 (0%)	1 (5%)	0 (0%)	1 (5%)	(1,88) = 2.1, p = 0.553

Table 9: Overarching Categories and Mid-Level Categories from Experiment 1 Applied to Cues of Human-Likeness.

overarching category	mid-level category	cue
predictable	observation of base behavior without deeper interpretation	interacted with environment
		skilled
		interacted with participant
		repeatitive
unpredictable	observation of base behavior without deeper interpretation	not enough interactions with participant
		not enough interactions with environment
		random/unpredictable behavior
		unskilled
		too much interaction with participant
		weird interactions with environment
	perception of ToM	learning
either	observation of base behavior without deeper interpretation	movement
		cooperative behavior
		competitive behavior

Game Theoretic Analysis

An empirical game theoretic analysis was conducted on participants' responses to how they would choose to interact strategically in the future. Results are presented in Table 10. As in Experiment 1, a coordination game, Stag Hunt and Deadlock were described by participants. In addition, participants also described Mutual Assured Destruction and a social dilemma (see Figure 1 for a description). Again, the ability to interact with the co-player may have affected participants' overall perceptions of co-players and resulted in more complex games being described.

Table 10: Empirical Game Theoretic Analysis of Responses from the DST Experiment. Five typical game theory games were derived from interview responses. As in Experiment 1, there was a coordination game, in which mutual competition was selected for fun, a Stag Hunt, in which mutual cooperation is selected for greater utility in achieving a higher score, and Deadlock, where the player chooses to compete regardless of what the other player decides because they believe they are more skilled and can win. In addition to the games seen in Experiment 1, analysis yielded Mutual Assured Destruction, in which players avoid conflict out of fear of the cost of retaliation, and a social dilemma, in which a player wants to engage in mutual cooperation, but has doubts about how much they can trust the second player, resulting in competition.

Participant's Game:	Coordination	Social dilemma	Mutual Assured Destruction	Stag Hunt	Deadlock
Example:	"Cooperatively because I prefer cooperative games than playing competitively" or "Competitive, because it is much more fun"	"I'd like to play cooperatively because we could live longer together, but based on interactions, not confident that there would be progress in the future through communicating. So I would compete"	"Cooperatively, because its the best way to stay safe" or "Don't want to compete with someone else since I was a beginner; and I think he would definately be willing to kill me"	"Cooperative because they helped me not starve and get through the night. This aided my survival"	"Competitive, because I can survive longer"
Confederate	7/22 (32%)	1/22 (5%)	5/22 (23%)	8/22 (36%)	0/22 (0%)
Tandem	10/22 (45%)	1/22 (5%)	2/22 (9%)	9/22 (41%)	0/22 (0%)
Simple AI	6/22 (27%)	0/22 (0%)	1/22 (5%)	10/22 (45%)	2/22 (9%)
Social AI	4/22 (18%)	8/22 (36%)	3/22 (14%)	7/22 (32%)	0/22 (0%)

Discussion

The same inverted U-shape is seen for perceived human-likeness by expertise as was seen in Experiment 1, though the peak is shifted to the right (see Figures 7 and 31). The reporting method varied between the two experiments, as did the percent of all agents that were perceived as humans, which likely affected these values. In general, again there is evidence that the intermediate level of competence has a higher association with perception of human-likeness than do extreme beginner or extreme expert levels.

Some cues (i.e., categories) used to determine human-likeness were similar to Experiment 1, including random behavior and repetitive or predictable behavior. Some examples include:

> Tandem Human as AI: *"The movement seemed a little more random than I expected."*
>
> Social AI as AI: *"It was very predictable when he would come up to me. The first thing I did to interact with player 2 was to give him something, and he attacked. If a person, it made sense if it was a person, but every other time he would come up and either give me something or attack."*
>
> Social AI as Human: *"Player 2 would attack for no reason; would give me stuff to help out, but biggest indicator was to protect, but [by the] 2nd night they were on the other side of the map doing something on their own."*

While some participants in this experiment often relied on expectations of humanlike interactions (resulting in predictability or unpredictability) to make determinations of human-likeness, there were overall many more participants who believed that the co-player was an AI, regardless of true identity, when compared to Experiment 1. This bias may have affected how perceptions of predictability and unpredictability were related to perceptions of human-likeness.

However, the most commonly used cues included whether or not the participant perceived that the co-player tried to interact with them, or if the co-player was interacting too much. Overall, 61.3%, SD = 44.5%, of participants based their decision of human-likeness in whole or in part on if and how the co-player interacted with them. It is likely that the ability to interact in Experiment 3 influenced participants' evaluation process and caused the cues/categories to be different than in Experiment 1.

From these results, the answer to research question 3 ("What characteristics of social interactions do humans use to differentiate between humans and AI, and how do Human-Human Interactions and Human-AI Interactions (HAI) differ within a complex environment?") seems to be that firstly, the way distinctions are made differs when the observation of behavior occurs with or without the opportunity to interact. In this experiment, even in encounters where interactions are not required but a matter of choice, participants often judged human-likeness based on whether or not, and how much, they perceived that a co-player tried to interact with them. Most often, human-likeness was assumed when the co-player was seen as being interactive, but not excessively interactive.

A number of differences between interactions with co-players were reported by participants. Most notably, while the Confederate Human was perceived as a human least often, he was liked and trusted more than any other co-player. Conversely, while the Social AI was perceived as a "real person" more often than any other player, and explicitly rated as a human second most frequently, it was trusted and liked less than any other player. This may indicate that the perception of human-likeness itself did not dictate the extent to which a co-player was liked or trusted, but the overall impression gained during the interaction

did. Additionally, it means that an entity need not be perceived as a human to be trusted and liked. While the qualitative results give some indications about what was different in the overall interactions with the different co-players, some of this may be related to the specific context of this experiment (i.e., playing a multiplayer videogame in a lab with a stranger of unknown identity without the ability verbally communicate). Still, it is interesting that while each participant was able to have a unique experience with the game and their co-player, some overall differences (like trusting, liking) varied significantly depending on co-players.

OVERALL DISCUSSION OF FINDINGS

While human perceptions of the behavior of non-human agents has been a topic of research for many years, the study of how complex, AI generated behaviors are perceived and what specific behavioral qualities lead to the perception of human-likeness has been missing.

Overall, the qualities of observable behaviors that affect the perception of human-likeness are complex. The pilot experiment (i.e., Experiment 1) hinted that perceived competence, predictability, and context may have an influence on the perception of human-likeness when observing behavior, and Experiments 2 and 3 explored these relationships more thoroughly.

Competence and Predictability

Interpretation of Findings

It makes intuitive sense that competence/skill would affect perceptions of humanness, as our typical experience with other people is that they can competently navigate the world around them, but AI is still not universally competent. From Experiment 1, as an AI becomes more competent, it seems to be perceived with increased perceptions of humanness. However, from the interviews, it appeared that maximum human-likeness was perceived at the level of intermediate, and that these perceptions decreased at the level of expert because humans may have expected that competent AI should never make any

mistakes (i.e., the Perfect Automation Schema) whereas competent humans still make mistakes for various reasons.

In Experiment 2, it was observed that average perceptions of humanness generally increased with average perceptions of skill/competence, but that perceived human-likeness was maximum at the level of expert and dipped lower at the level of intermediate. The reason for this seemed to be due to a high degree of variability in perceived human-likeness. This result, as well as the cluster analysis of character traits, seemed to show no evidence that the Perfect Automation Schema affected perceptions of human-likeness in this study.

The relationship between perceived competence, predictability and human-likeness was also important to further investigate. In Experiment 1, while competence and humanness were intentionally manipulated, predictability was not directly manipulated, yet the sentiment came up often in interviews. Participants' statements and previous literature seemed to indicate a relationship in which behaviors that were unpredictable could be perceived as human-like due to a desire to better understand the player (i.e., the effectance motivation; Waytz et al., 2010b), but the perception that behaviors were predictable could either come from the perception that they were simple and repetitive (more often associated with a beginner level of skill), or that they were predictable because the player had human-like thoughts and motivations that produced the behavior (e.g., having thoughts or reasoning, experimenting with the level, or having goals; ToM states), and these ToM states were more often associated with the perception of humanness and competence.

In Experiment 2, it was determined that relationship between average perceived humanness and average perceived predictability was non-linear, and that average perceived

predictability differed across 3 levels of average perceived skill (main effect of skill level and nearly an interaction between skill level and the quadratic contrast). In particular, at the level of expert, the relationship between perceived humanness and skill had the maximal U-shape, where perceived humanness was high at both the high and low ends of perceived predictability. There is also some evidence from the cluster analysis that ToM was perceived when human-likeness and skill were sufficiently high, as the cluster containing the majority of ToM states had significantly higher average perceived humanness, skill and predictability. This possibly provides support to the theory that at sufficiently high levels of skill, performances can be perceived as predictable and humanlike due to the ability to perceive them as the result of humanlike thoughts and motivations (i.e., applying ToM).

Significance of Results

While the relationship between competence and human-likeness has previously been assumed in the literature (Fisk, Cuddy & Glick, 2006; Waytz et al., 2010a), Experiment 2 goes further to investigate that connection than any known currently published research. While Waytz, Heafner & Epley (2014) were able to show that increased human-likeness in an autonomous vehicle could increase the perception of competence, human-likeness was manipulated through adding humanlike features like a name, gender and voice, and competence was not manipulated at all. In this study, all physical features stayed constant across all players (all used the avatar Mario and no sound was included), and the only difference was how players behaved within the game. By having participants observe these

behavioral differences across players, it was seen that perceived humanness did increase with perceived competence (i.e., skill). However, the causal relationship was not determined by this study (i.e., does an increase in perceived human-likeness increase the perception of competence, does the perception of competence increase the perception of human-likeness, or do both affect the perception of the other).

Interestingly, in Experiment 2 overall, average perceived humanness increased with predictability (i.e., decreased with unpredictability). This was contrary to the hypothesis for the experiment as well as prior literature (Waytz et al., 2010b; Short et al., 2010; Hayes et al., 2014; Salem et al., 2013; Kompatsari et al. 2019). It is possible that this is related to the fact that the stimuli was overall perceived with higher levels of human-likeness and skill. If this was the case, it is also possible that if stimuli was included for players (human and AI) that had lower skill, the pattern of perceived human-likeness and predictability for low skilled players would have more closely matched what was predicted. However, it is also possible that the results hold true and that in the context of playing a videogame, especially one that is so well known throughout the world, people associate unpredictability at a lower skill level with non-humanness. This is actually in line with the findings of Pfeiffer et al. (2011), in which, (1) context affected the relationship between predictability and perceived humanness, and (2) that in certain contexts, maximum unpredictability had the lowest levels of perceived human-likeness.

Still, Experiment 2 has demonstrated that the relationship of perceived human-likeness and predictability of actions is affected by the perceived level of skill, and that at high levels of skill, low and high values of predictability are associated with high values

of perceived human-likeness. To my knowledge, no other research has specifically explored this relationship before. This may indicate that at sufficiently high levels of perceived skill and human-likeness, actions are interpreted as the result of a human-like mind with thoughts and motivations (i.e., applying ToM), and that this provides a lens to better explain behavior.

Context of Interaction and Social Context

After the pilot study, what remained unclear was if the same qualities used to make determinations of humanness solely on the basis of observed behavior would be the same qualities used in a context more related to real world interactions, where participants could decide to interact with the player however they wanted when making determinations of humanness. From the empirical game theoretic analysis, there was some indication that social context was perceived differently for human and AI players. However, in Experiment 1, participants were asked at the end to select any agent to play a cooperative or competitive game with, and so there was no evidence of how they would respond for individual players if they did not know their true identity.

Interpretation of Findings

Overall, whether or not, and how co-players engaged participants in social interactions seemed to be the cue most commonly used by participants to determine humanness, though it was not always a reliable cue of true humanness. In the context of a less ecologically

valid experiment, or experiments that do not include the potential for social interaction (e.g., Experiments 1 and 2), this finding would not be possible.

Additionally, competence was again manipulated in Experiment 3, but the manipulation was slightly different. Experiments 1 and 2, competence was manipulated by including "beginner" and "expert" human and AI players (i.e., varying skill). In Experiment 3, competence was varied for the human players by including a Confederate Human player who became skilled over the course of the experiment (i.e., competent or skilled) and including participants as co-players for other participants in the tandem condition who had never played the game before (i.e., unskilled). For the AI co-players, the Simple AI had low competence in the sense that it had a simple behavior tree that was only intended to give the appearance that it was playing the game, when in reality, it paid no attention to its own performance (i.e., health, hunger, sanity) or the participant in the game. The Social AI, on the other hand, had competence in the sense that it actually attempted to play the game and interact in social ways with the participant (i.e., the cognitively plausible behavior tree for survival behavior and POMDP to decide social behavior). In the context of the interactive game, the increase in competence of the Social AI did seem to be accompanied with an increase in the perception of human-likeness, but the arguably more "competent" human confederate player received the lowest explicit ratings of humanness and lower ratings of perceived human-likeness (i.e., "real person" and the GS questionnaires) than the Tandem Human players. It is possible that this is related to the fact that the human confederate had the largest average in-game distance between players (i.e., he stayed further away on average than did the other co-players),

making it difficult for players to interact with him over the course of the experiment. Since engaging in social interactions was very commonly used as a cue to indicate humanness, is possible that this distance and lack of interaction caused the Confederate Human to be perceived as less human-like. Another possibility is that this perception was related to his perceived competence. The Confederate Human was perceived as an expert most often, and it is possible that something like the Perfect Automation Schema caused him to be perceived as more AI-like (i.e., he's too good to be human). The perception that he was an expert may also contribute to positive perceptions of trust and liking, though this could not be said conclusively from these results.

Additionally, the social context described in interviews and coded using the empirical game theoretic analysis differed from Experiments 1 and 2. GT games described in Experiment 1 were further explored in Experiments 2 and 3. In Experiment 3, while the same 3 GT games found in Experiment 1 were also found in Experiment 3, 2 additional games (Mutual Assured Destruction and a social dilemma). It seems likely that the context of Experiment 3 (i.e., an interactive game) influenced the overall perceived relationships participants developed with co-players. Afterall, relationships are developed over time and as a result of how we interact with others. In Experiment 3, participants had much more time to observe the co-player, and were actually able to interact with them unlike in Experiments 1 and 2. While Experiment 2 included all 5 games (which was translated to 6 games when the coordination game was split into cooperative or competitive preferences), the results more closely resembled those of Experiment 1, again offering some indication that allowing for interactions in a complex environment rather than simply observing brief

vignettes of behavior can affect the overall perception of social context with an interaction partner.

Significance of Results

Results from Experiment 3 show that social interaction is an expectation of human-likeness. Overall, humans were not very sensitive to true humanness in the game environment when they could only interact and observe non-verbally. Overall, all players were rated humans less than 50% of the time, however, there were differences between the co-players that speaks to the overall effect of competence on perceived human-likeness. The pattern in Figure 31 is the same as is seen in Figure 7 (from Experiment 1), though this pattern is not present in Experiment 2. While the reason for this pattern is not totally clarified by these three experiments, it speaks to the fact that the perception of human-likeness based on the observation of behavior is complex, and the evaluation is affected by context as well as perceived competence. Future research would be necessary to fully clarify the relationship between competence and human-likeness, as well as identify other factors that influence perceptions of human-likeness at middle and high levels of perceived competence.

Additionally, the Turing Test is still a standard for evaluating the human-likeness of AI. While in the original Turing Test, participants only have 5 minutes of unrestricted conversation via text, Experiment 3 allowed participants unrestricted interaction for 30 minutes within a complex and dynamic environment, and the Social AI still passed the Turing Test (i.e., greater than 30% perceived the co-player to be a human). This co-player

was designed to be competent and human-like by using a cognitively plausible behavior tree for survival behaviors as well as a learning algorithm to interact socially with participants. In other words, the Social AI co-player was developed with a human-like mind that decided for itself how to play the game and interact, and this did lead to greater perceptions of human-likeness over a Simple AI. However, while the Social AI was perceived as a human more frequently, it had the lowest levels of perceived trust and liking. This finding is in line with previous research that shows that an increase in the perception of human-likeness of a non-human is often accompanied by consequences in how positively they are perceived overall (Waytz et al., 2010b; Short et al., 2010; Hayes et al., 2014). Interestingly, the Confederate Human player received low ratings of human-likeness and was explicitly identified as a human least often, yet received the highest ratings of skill, trust and liking. This is a hopeful outcome for future HAI, as it indicates that the perception of human-likeness is not required for a co-player to be trusted and liked, although it also suggests that there are still qualities of true humanness that need to be understood and incorporated into an AI to lead to these positive perceptions.

REFERENCES

Axelrod, R., & Hamilton, W. D. (1981). The evolution of cooperation. *Science*, 211(4489), 1390-1396. http://www.psych.ualberta.ca/~phurd/cruft/Axelrod-science.pdf

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a "theory of mind". *Cognition*, 21(1), 37-46.

Bartneck, C., Kulić, D., Croft, E., & Zoghbi, S. (2009). Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. International journal of social robotics, 1(1), 71-81.

Bisio, A., Sciutti, A., Nori, F., Metta, G., Fadiga, L., Sandini, G., & Pozzo, T. (2014). Motor contagion during human-human and human-robot interaction. *PloS One*, 9(8).

Blakemore, S. J., & Decety, J. (2001). From the perception of action to the understanding of intention. *Nature Reviews Neuroscience*, 2(8), 561-567. https://www.nature.com/articles/35086023

Bling, S. [SethBling]. (2015, June 13) *MarI/O-machine learning for video games.* [Video] Youtube. https://www.youtube.com/watch?v=qv6UVOQ0F44

Breazeal, C., & Scassellati, B. (1999, October). How to build robots that make friends and influence people. *IEEE/RSJ International Conference on Intelligent Robots and Systems*. (Vol. 2, pp. 858-863). http://cs-www.cs.yale.edu/homes/scaz/papers/Breazeal-Scaz-IROS99.pdf

Caruana, N., de Lissa, P., & McArthur, G. (2017). Beliefs about human agency influence the neural processing of gaze during joint attention. *Social Neuroscience*, 12(2), 194-206. https://www.tandfonline.com/doi/full/10.1080/17470919.2016.1160953

Castelli, F., Happé, F., Frith, U., & Frith, C. (2000). Movement and mind: a functional imaging study of perception and interpretation of complex intentional movement patterns. *Neuroimage*, 12(3), 314-325.

Dennett, D. C. (1989). The Intentional Stance. MIT Press.

Dworkin, S.L. Sample Size Policy for Qualitative Studies Using In-Depth Interviews. Arch Sex Behav 41, 1319–1320 (2012). https://doi.org/10.1007/s10508-012-0016-6

Dzindolet, M. T., Pierce, L. G., Beck, H. P., Dawe, L. A. (2002). The perceived utility of human and automated aids in a visual detection task. *Human Factors*, 44, 79–97.

Ehsan, U., Harrison, B., Chan, L., & Riedl, M. O. (2018, December). Rationalization: A neural machine translation approach to generating natural language explanations. *In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 81-87). ACM. https://dl-acm-org.mutex.gmu.edu/citation.cfm?id=3278736

Emami, P., Hamlet, A. J., & Crane, C. D. (2015) POMDPy: An Extensible Framework for Implementing Partially-Observable Markov Decision Processes in Python.

Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: a three-factor theory of anthropomorphism. *Psychological Review*, 114(4), 864.

Epstein, R., Roberts, G., & Beber, G. (Eds.). (2009). Parsing the Turing Test. Springer Netherlands.

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior research methods*, 39(2), 175-191.

Fiske, S. T., Cuddy, A. J., & Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. *Trends in Cognitive Sciences*, 11(2), 77-83. http://www.cogsci.bme.hu/~ktkuser/kepek/journalreadingclub2010/13fiskesoccog.pdf

Frischen, A., Bayliss, A. P., & Tipper, S. P. (2007). Gaze cueing of attention: visual attention, social cognition, and individual differences. *Psychological Bulletin*, 133(4), 694.

Frith, C. D., & Frith, U. (1999). Interacting minds--a biological basis. *Science*, 286(5445), 1692-1695. https://science.sciencemag.org/content/286/5445/1692

Gazzola, V., Rizzolatti, G., Wicker, B., & Keysers, C. (2007). The anthropomorphic brain: the mirror neuron system responds to human and robotic actions. *Neuroimage*, 35(4), 1674-1684.

Glaser, B. G. (1998). Doing Grounded Theory. Mill Valley, CA: Sociology Press.

Glaser, B. G., & Strauss, A. L. (1967). The Discovery of Grounded Theory: Strategies for Qualitative Research. New York, NY: Aldine

Gouldner, A. W. (1960). The norm of reciprocity: A preliminary statement. *American Sociological Review*, 161-178. https://www-jstor-org.mutex.gmu.edu/stable/2092623?seq=1#metadata_info_tab_contents

Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of mind perception. *Science,* 315(5812), 619-619.

Hadfield-Menell, D., Russell, S. J., Abbeel, P., & Dragan, A. (2016). Cooperative inverse reinforcement learning. In Advances in Neural Information Processing Systems (pp. 3909-3917).

Hayes, B., Ullman, D., Alexander, E., Bank, C., & Scassellati, B. (2014, August). People help robots who help others, not robots who help themselves. *In The 23rd IEEE International Symposium on Robot and Human Interactive Communication* (pp. 255-260). IEEE. http://www.bradhayes.info/papers/roman14.pdf

Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *The American Journal of Psychology*, 57(2), 243-259. http://cs.engr.uky.edu/~sgware/reading/papers/heider1944experimental.pdf

Huang, R. (2014). RQDA: R-based qualitative data analysis. R package version 0.2–7.

Kompatsiari, K., Ciardo, F., Tikhanoff, V., Metta, G., & Wykowska, A. (2019). It's in the Eyes: The Engaging Role of Eye Contact in HRI. *International Journal of Social Robotics*, 1-11. https://link-springer-com.mutex.gmu.edu/article/10.1007/s12369-019-00565-4

Laird, J., & VanLent, M. (2001). Human-level AI's killer application: Interactive computer games. *AI Magazine,* 22(2), 15-15.

Lee, D. (2008). Game theory and neural basis of social decision making. *Nature Neuroscience*, 11(4), 404-409. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2413175/

Lee, K. W., & Hwang, J. H. (2008). Human–robot interaction as a cooperative game. *In Trends in Intelligent Systems and Computer Engineering* (pp. 91-103). Springer, Boston, MA.

Leibo, J. Z., Zambaldi, V., Lanctot, M., Marecki, J., & Graepel, T. (2017). Multi-agent reinforcement learning in sequential social dilemmas. *arXiv preprint arXiv:1702.03037*.

Lim, S., & Reeves, B. (2009). Being in the game: Effects of avatar choice and point of view on psychophysiological responses during play. *Media Psychology*, 12(4), 348-370.

MacDorman, K. F. (2006, July). Subjective ratings of robot video vignettes for human likeness, familiarity, and eeriness: An exploration of the uncanny valley. *In*

ICCS/CogSci-2006 long symposium: Toward social mechanisms of android science (pp. 26-29).

MacDorman, K. F., & Ishiguro, H. (2006). The uncanny advantage of using androids in cognitive and social science research. *Interaction Studies*, 7(3), 297-337.

Martin, A., & Weisberg, J. (2003). Neural foundations for understanding social and mechanical concepts. *Cognitive Neuropsychology*, 20(3-6), 575-587. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1450338/pdf/nihms9563.pdf

McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., ... & Etemadi, M. (2020). International evaluation of an AI system for breast cancer screening. *Nature*, 577(7788), 89-94.

McNamara, J. M., Barta, Z., Fromhage, L., & Houston, A. I. (2008). The coevolution of choosiness and cooperation. *Nature*, 451(7175), 189-192.

Merritt, S. M., Unnerstall, J. L., Lee, D., & Huber, K. (2015). Measuring individual differences in the perfect automation schema. *Human Factors*, 57(5), 740-753.

Miller, J. H. (1996). The coevolution of automata in the repeated prisoner's dilemma. *Journal of Economic Behavior & Organization*, 29(1), 87-112.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., ... & Petersen, S. (2015). Human-level control through deep reinforcement learning. *Nature*, 518(7540), 529-533.

Mutlu, B., Yamaoka, F., Kanda, T., Ishiguro, H., & Hagita, N. (2009, March). Nonverbal leakage in robots: communication of intentions through seemingly unintentional behavior. *In Proceedings of the 4th ACM/IEEE International Conference on Human Robot Interaction* (pp. 69-76). https://dl-acm-org.mutex.gmu.edu/citation.cfm?id=1514110

Osawa, H., Tobita, K., Kuwayama, Y., Imai, M., & Yamada, S. (2012, September). Behavioral Turing test using two-axis actuators. *In 2012 IEEE RO-MAN: The 21st IEEE International Symposium on Robot and Human Interactive Communication* (pp. 328-333). IEEE.

Palaniappan, M., Malik, D., Hadfield-Menell, D., Dragan, A., & Russell, S. (2017). Efficient cooperative inverse reinforcement learning. *Proc. ICML Workshop on Reliable Machine Learning in the Wild*.

Pfeiffer, U. J., Timmermans, B., Bente, G., Vogeley, K., & Schilbach, L. (2011). A non-verbal turing test: differentiating mind from machine in gaze-based social

interaction. *PloS One*, 6(11), e27591. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0027591

Poulin-Dubois, D., Crivello, C., & Wright, K. (2015). Biological motion primes the animate/inanimate distinction in infancy. *PloS One*, 10(2). https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0116910

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

Rabinowitz, N. C., Perbet, F., Song, H. F., Zhang, C., Eslami, S. M., & Botvinick, M. (2018). Machine theory of mind. *arXiv preprint arXiv:1802.07740.* https://arxiv.org/abs/1802.07740

Rakison, D. H., & Poulin-Dubois, D. (2001). Developmental origin of the animate–inanimate distinction. *Psychological Bulletin*, 127(2), 209. https://psycnet-apa-org.mutex.gmu.edu/fulltext/2001-16969-002.pdf

Ross, D. (2001). Stanford Encyclopedia of Philosophy: Game Theory. Disponibile da https://plato.stanford.edu/entries/game-theory.

Salem, M., Eyssel, F., Rohlfing, K., Kopp, S., & Joublin, F. (2013). To err is human (-like): Effects of robot gesture on perceived anthropomorphism and likability. *International Journal of Social Robotics*, 5(3), 313-323. https://link-springer-com.mutex.gmu.edu/article/10.1007/s12369-013-0196-9

Sally, D. (1995). Conversation and cooperation in social dilemmas: A meta-analysis of experiments from 1958 to 1992. *Rationality and Society*, 7(1), 58-92.

Sandoval, E. B., Brandstetter, J., Obaid, M., & Bartneck, C. (2016). Reciprocity in human-robot interaction: a quantitative approach through the prisoner's dilemma and the ultimatum game. *International Journal of Social Robotics*, 8(2), 303-317. https://link.springer.com/article/10.1007/s12369-015-0323-x

Sanfey, A. G., Rilling, J. K., Aronson, J. A., Nystrom, L. E., & Cohen, J. D. (2003). The neural basis of economic decision-making in the ultimatum game. *Science*, 300(5626), 1755-1758.

Scassellati, B. (2002). Theory of mind for a humanoid robot. *Autonomous Robots*, 12(1), 13-24. http://groups.csail.mit.edu/lbr/hrg/2000/Humanoids2000-tom.pdf

Schneider, E. F., Lang, A., Shin, M., & Bradley, S. D. (2004). Death with a story: How story impacts emotional, motivational, and physiological responses to first-person shooter video games. Human communication research, 30(3), 361-375.

Shigemi, S., Goswami, A., & Vadakkepat, P. (2019). ASIMO and humanoid robot research at Honda. *In Humanoid robotics: A reference* (pp. 55-90). Springer.

Short, E., Hart, J., Vu, M., & Scassellati, B. (2010, March). No fair!! an interaction with a cheating robot. *In 2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 219-226). IEEE. https://ieeexplore.ieee.org/abstract/document/5453193

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., ... & Chen, Y. (2017). Mastering the game of go without human knowledge. *Nature*, 550(7676), 354-359.

Stanley, K. O., & Miikkulainen, R. (2002). Evolving neural networks through augmenting topologies. *Evolutionary Computation*, 10(2), 99-127.

Suter, W. N. (2012). Qualitative data, analysis, and design. Introduction to Educational Research: A Critical Thinking Approach, 2, 342-86.

Syed, M., & Nelson, S. C. (2015). Guidelines for establishing reliability when coding narrative data. *Emerging Adulthood*, 3(6), 375-387.

Takahashi, H., Terada, K., Morita, T., Suzuki, S., Haji, T., Kozima, H., ... & Naito, E. (2014). Different impressions of other agents obtained through social interaction uniquely modulate dorsal and ventral pathway activities in the social human brain. *Cortex*, 58, 289-300. https://www-sciencedirect-com.mutex.gmu.edu/science/article/pii/S0010945214001142

Trivers, R. L. (1971). The evolution of reciprocal altruism. *The Quarterly Review of Biology*, 46(1), 35-57.

Tulk, S., Cumings, R., Zafar, T., & Wiese, E. (2018). Better know who you are starving with: Judging human-likeness in a multiplayer videogame. *In Proceedings of the Technology, Mind, and Society* (pp. 1-6).

Turing, M. (1950). Computing machinery and intelligence. *Mind*, 59(236), 433.

Vinyals, O., Babuschkin, I., Czarnecki, W. M., Mathieu, M., Dudzik, A., Chung, J., ... & Oh, J. (2019). Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*, 575(7782), 350-354.

Wang, Y., & Quadflieg, S. (2015). In our own image? Emotional and neural processing differences when observing human–human vs human–robot interactions. *Social Cognitive and Affective Neuroscience*, 10(11), 1515-1524.

Waytz, A., Gray, K., Epley, N., & Wegner, D. M. (2010a). Causes and consequences of mind perception. *Trends in Cognitive Sciences*, 14(8), 383-388. https://www.ncbi.nlm.nih.gov/pubmed/20579932

Waytz, A., Heafner, J., & Epley, N. (2014). The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. *Journal of Experimental Social Psychology*, 52, 113-117.

Waytz, A., Morewedge, C. K., Epley, N., Monteleone, G., Gao, J. H., & Cacioppo, J. T. (2010b). Making sense by making sentient: effectance motivation increases anthropomorphism. *Journal of Personality and Social Psychology*, 99(3), 410. https://pdfs.semanticscholar.org/3713/ad51a1900f524b91b6184760709dfec8adad.pdf

Wellman, M. P. (2006, July). Methods for empirical game-theoretic analysis. *In AAAI* (pp. 1552-1556).

Wiese, E., Metta, G., & Wykowska, A. (2017). Robots as intentional agents: using neuroscientific methods to make robots appear more social. *Frontiers in Psychology*, 8, 1663. https://www.frontiersin.org/articles/10.3389/fpsyg.2017.01663/full

Wykowska, A., Kajopoulos, J., Obando-Leitón, M., Chauhan, S. S., Cabibihan, J. J., & Cheng, G. (2015). Humans are well tuned to detecting agents among non-agents: examining the sensitivity of human perception to behavioral characteristics of intentional systems. *International Journal of Social Robotics*, 7(5), 767-781. https://link-springer-com.mutex.gmu.edu/article/10.1007/s12369-015-0299-6

Wykowska, A., Wiese, E., Prosser, A., & Müller, H. J. (2014). Beliefs about the minds of others influence how we process sensory information. *PLoS One*, 9(4), e94339.

BIOGRAPHY

Stephanie Tulk Jesso received her Bachelor of Science in civil engineering from Michigan Technological University in 2012. She did some soul searching for 2 years before switching careers to human factors, and received her Master of Arts in psychology with a concentration in human factors and applied cognition from George Mason University in 2016. She currently works as a human factors consultant at Carilion Clinic in Roanoke VA.