Effects of Communication Directionality and AI Agent Differences in Human-AI Interaction

Zahra Ashktorab
IBM Research AI
NY, USA
zahra.ashktorab1@ibm.com

Casey Dugan
IBM Research AI
NY, USA
cadugan@us.ibm.com

James Johnson
IBM Research AI
Cambridge, Massachusetts, USA
jmjohnson@us.ibm.com

Qian Pan
IBM Research AI
Cambridge, Massachusetts, USA
qian.pan@us.ibm.com

Wei Zhang
IBM Research AI
NY, USA
zhangwei@us.ibm.com

Sadhana Kumaravel
IBM Research AI
NY, USA
sadhana.kumaravel1@us.ibm.com

Murray Campbell
IBM Research AI
NY, USA
mcam@us.ibm.com

ABSTRACT

In Human-AI collaborative settings that are inherently interactive, direction of communication plays a role in how users perceive their AI partners. In an AI-driven cooperative game with partially observable information, players (be it the AI or the human player) require their actions to be interpreted accurately by the other player to yield a successful outcome. In this paper, we investigate social perceptions of AI agents with various directions of communication in a cooperative game setting. We measure subjective social perceptions (rapport, intelligence, and likeability) of participants towards their partners when participants believe they are playing with an AI or with a human and the nature of the communication (responsiveness and leading roles). We ran a large scale study on Mechanical Turk (n=199) of this collaborative game and find significant differences in gameplay outcome and social perception across different AI agents, different directions of communication and when the agent is perceived to be an AI/Human. We find that the bias against the AI that has been demonstrated in prior studies varies with the direction of the communication and with the AI agent.

CCS CONCEPTS

• **Human-centered computing** → *Collaborative interaction*; **Empirical studies in collaborative and social computing**.

KEYWORDS

human-AI interaction, games, collaboration, social perception

CHI '21, May 8–13, 2021, Yokohama, Japan

ACM ISBN 978-1-4503-8096-6/21/05…$15.00
https://doi.org/10.1145/3411764.3445256

ACM Reference Format:
Zahra Ashktorab, Casey Dugan, James Johnson, Qian Pan, Wei Zhang, Sadhana Kumaravel, and Murray Campbell. 2021. Effects of Communication Directionality and AI Agent Differences in Human-AI Interaction. In *CHI Conference on Human Factors in Computing Systems (CHI '21), May 8–13, 2021, Yokohama, Japan.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3411764.3445256

1 INTRODUCTION

Interactivity, which includes responsiveness [53] and direction of communication [38] has been an important aspect of human-computer interaction as it can impact both user experience and social perception. Given the pervasiveness of AI in collaborative spaces, many HCI researchers are investigating factors that lead to successful outcomes of these collaborative exchanges [24, 41, 55]. Researchers have also investigated user perceptions of AIs in these spaces [50, 62], since social perception of one's partner in a collaborative space can impact the outcome of the collaboration. The direction of communication (i.e. who is "leading" the interaction) in Human-AI collaborative spaces is important to study given that AI is being used increasingly in collaborative spaces, including chatbots being employed for tutoring [23], collaborative writing [14] or even as virtual nurse agents in hospitals [9], and how an update to increase AI performance may even hurt team performance [5].

Cooperative partially observable games (CPO) [33] are cooperative games in which there is information hidden from partners. These games pose challenges to AI researchers as they require researchers to consider theory of mind [7]. While in the past there have been many strides made in competitive zero-sum games (chess, poker, checkers) [10, 13], cooperative partially observable games require users to interact with their partners in a way that the information being communicated is interpretable (without giving away all of the information), and also interpreting partially visible information. One example of such a game that has been widely studied is Hanabi [7]. In particular, CPOs, with their focus on interpreting partial information and game dynamics which can include

turn-taking, naturally lend themselves to studying directionality of communication between AI agents and humans.

In this paper, we investigate interactions in an AI-driven cooperative word game that requires players to play one of two roles: 1) interpret an AI agent's clues or 2) send interpretable clues to the AI agent to help their partners guess a target word. Prior work investigating social perceptions of AI agents in a cooperative game with partially observable information has focused on one direction of communication, with human player acting as "guesser", i.e. the AI agent giving clues to the human player and the player attempting to guess the target word [4]. In this work, we investigate both directions of communication ("giver" and "guesser") with various AI agents that have been trained differently and behave/interact differently. "Guesser" and "giver" agent roles behave differently in responsiveness and direction of communication. When an individual plays as the "giver" player, they are providing clues to the AI agent, "steering" it in the correct direction. When the user plays as the "guesser" player, the AI does very little adjusting to steer the individual. Based on these characteristics, it is fair to hypothesize that users will find the agents more likeable when they play the role of "giver" since this role provides more control and agency to the user. We consider these differences as we investigate the social perceptions of the AI agents in this study. Prior work has also investigated social perceptions in the context of one AI agent, with one knowledge base. In this paper, we explore social perceptions and interactions with multiple agents to identify if agents that are trained differently have an impact on social perceptions in the context of directionality and AI identity perception.

We investigate the following research questions:

> **RQ1** What are the social perceptions of an individual's partner in a AI-driven cooperative game with partially observable information?
>
> How does the **direction of the communication**, **various AI agents**, and **perception of their partner's AI identity** affect these social perceptions?
>
> **RQ2** How is the outcome of collaboration impacted in a AI-driven cooperative game with partially observable information?
>
> How does the **direction of the communication**, **various AI agents**, and **perception of their partner's AI identity** affect the outcome of the collaboration?

In this paper, we present an online study (n=199) in which participants played *Guess the Word*, a word guessing game similar to [3, 21], with multiple AI agents in various directions of communication (as "guesser" and "giver") and filled out a survey about their social perceptions. We show that there are social perception differences with different AI agents, as there are when the direction of communication varies. We show that the bias against the AI that has been demonstrated in prior studies [4, 20] varies with the direction of communication and with the AI agent and that performance differences vary depending on how people interact with their perceived AI partner given the direction of communication.

2 RELATED WORK

Our inquiry is motivated by the widespread use of AI systems in human-AI collaborative environments. In this work, we build on prior work that investigates human-AI interaction [4, 9, 23] by investigating human-AI collaboration in various contexts including varying the direction of the communication in the collaboration and interaction with different AI agents.

2.1 Interactivity: Responsiveness, Direction of Communication, and User Control

Perceived interactivity was originally based on a two dimensional construct which includes (1) a users' psychological sense of efficacy and (2) users' sense of a system's interactivity [45]. This means that a system that is perceived as interactive can take user input and can execute on it [45, 60]. Additional research has identified an additional construct of perceived interactivity: "direction of communication", or the belief that two-way communication exists in the system [38]. The perceived interactivity in systems - or lack thereof - can impact the user experience and social perceptions individuals have of the system and the parties with which they are communicating [62]. Interactivity, which includes responsiveness [53], two-way communication [38], and user control [19] has been an important aspect of Computer-Mediated Communication and impacts social perceptions of individuals through the web [50]. Many individuals see interactivity as encompassing a two-way communication in which a user can play both the sender and receiver roles [19]. In this paper, we investigate interactivity, more specifically the directionality of communication by comparing the "giver" and "guesser" roles played by the user.

2.2 Cooperative Games with Partially Observable Information

There have been many AI-infused word association games developed, resulting in studies on how users understand, interact and communicate in these games [67]. Games like the 'Taboo' word game [54] "forces agents to speculate about their partner's understanding of the domain, rather than just performing inference on their own knowledge," as do games like "Hanabi" [6]. There has been work in which AI agents have been trained to play these games as a way of testing theories around how people interact and communicate, but ultimately as a contribution to furthering Artificial Intelligence research [1, 15, 25]. Other research has investigated communication in games like 'Password' and found that people playing both roles (speakers and hearers) are collaborative and considerate of one another [68]. Liang et al. investigated implicature communication in Hanabi, a cooperative partially observable game [37] that has been studied in AI literature as a CPO (cooperative game with partially observable information) [35]. In their study, they found that an Implicature AI, i.e. one that implicitly communicates information, led to a more successful outcome than non-implicature AIs.

2.3 Human-AI Collaboration

The term "human-AI collaboration" has emerged to describe interactions between AI systems and humans [2, 12, 49, 65]. Prior work has investigated this kind of collaboration in various domains including drawing pictures collaboratively [49], collaborating with AI systems to automate data science projects [65]. Both of these studies showed that while they were challenges, users remained positive

about the collaboration with their AI counterparts. In this paper, we study human-AI collaboration in the context of a game. Games are frequently used as a test bed for state-of-the-art AI algorithms and an important application domain of AI [13, 22, 58, 64].

One aspect of in collaborative games is the social perception of one's partner. Prior work has found that in human-AI collaborative games, when users perceive their partners to be human, they find them to be more intelligent and likeable [3]. In this study, we build on this work, by exploring the different roles in human-AI collaborative games (interpreting your partner's actions and sending digestible interpretable signals). We also investigate the interactions given AI agents that behave differently.

A few recent works explored the effect of disclosure. Shi et al. found that people are less persuaded by an AI chatbot even when the same dialogues are used by human interlocutors [57]. Focusing on cooperative behavior in a repeated prisoner's dilemma game, Ishowo-Oloko et al. found that disclosing the bot nature averts people's tendency to cooperate, and participants do not recover despite experiencing cooperative attitudes exhibited by bots [29]. These works show that it is not the displayed identity but the perceived identity that impacts the outcome, as people still suspected the identity of the agent, despite the display.

We expand prior work [4] that investigates the effect of disclosure of AI-identity for an AI partner that has human-comparable performance, on people's social perception and performance in one particular role for one particular agent. In this work, we investigate social perception of the AI agents in the various roles played in a human-AI collaborative context and with various AI agents. Ashktorab et al. found that users who believe that they are interacting with a human (when told that they are interacting with a human) have higher regard for their partner (i.e. find their partners more intelligent, likeable, create, have more rapport), whereas when they believe they are interacting with an AI they find their partners less intelligent. That study investigated one role played by the human in the human-AI collaboration, the "guesser", or the individual who has information withheld from them as they collaborate with their AI agent and responds with guesses to the AI agent's hints. That work only studied this collaboration within the context of one AI agent. We expand that study by investigating all roles ("giver" and "guesser") in which there are varying directions of the communication and responsiveness between the human and the AI as well as investigating interactions with multiple AI agents with varying knowledge bases.

2.4 Comparing "Performance" for Multiple AI Agents

Research comparing *performance* of AI agents has historically been published at AI-centric conference in which metrics like precision, recall, AUC and other metrics are of importance, and not necessarily the user reaction to those agents. On Reinforcement Learning for games and dialogues, [16, 18, 25, 34] use multi-agent RL to learn cooperative games with images as secret information. Li et al. [36] learns multi-agent dialogue agents by optimizing a shared objective that encourages dialogue flow. He et al. [26] trains symmetric dialogue agents and Hu et al. [28] uses reinforcement learning to learn agents that play 20 questions game against a static "guesser" simulator, where the multi-turn strategy was modeled as a Markov decision process. For CHI Audiences, studies often take one AI system and evaluate user reactions to one aspect of one AI system, whether it be conversational agents [48], computer mediated communication [30] or other forms of human-AI interaction [40]. In this work, we introduce six AI agents and find differences in the user perceptions to these various agents showing that it is simply not enough to study human-AI interaction in the context of one AI agent.

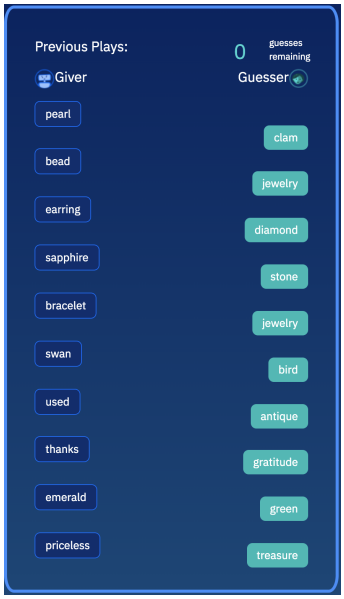
3 AI AGENT DESCRIPTION: A COOPERATIVE WORD GUESSING GAME

To learn about user perceptions of their opponent in a collaborative setting, we used a simple two-person collaborative game we call *Guess the Word*, similar to Wordgame in [4] and 'Passcode' [21] in prior work. In *Guess the Word*, the opponent has a target word and gives clues to their partner so that their partner guesses the target word. We refer to the player who is giving hints as the "giver" and the player who is guessing as the "guesser". The game begins with the AI starting the game with a hint like "car". After every hint, the player inputs a guess. In this example, the target word is "engine". Players get 10 attempts to guess before they lose. If players input the correct target word, they win. Figure 1 shows what a round with the different AI agents looks like. *Guess the Word* is cooperative, meaning players work together for the "guesser" to correctly guess the secret target word based on the hints provided by the "giver". The cooperative nature of this game means that players are open and honest in achieving a shared goal. In this section, we present the technical aspects of the models (Model A, Model B, and Model C) followed by their characteristics. Each of these models consists of two agents that play with users as the "giver" agent or the "guesser" agent. In essence, there are six agents, a "giver" and "guesser" for Model A, Model B and Model C. All references to the models in the document to reflect a human-centered perspective i.e Model A ("guesser") means that the user interacts with Model A as the "guesser". All AI agents are trained separately.

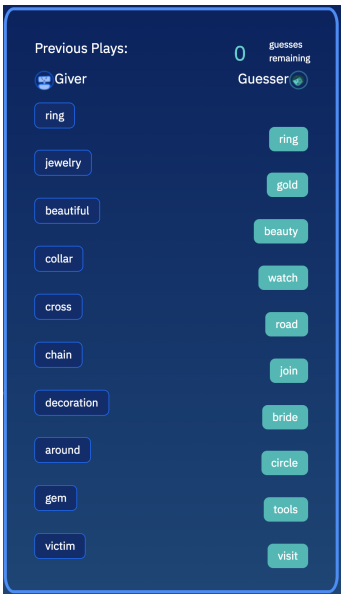
3.1 Model A: Supervised Learning

Model A agents use the target word to generate candidates using Free Association Norm to get the corresponding words (as clues) that lead to the secret word or all the secret words (as guesses) that could lead from a clue, word embeddings to get the top-k most similar words to the secret word or clue by cosine similarity, and WordNet [52] to get all the related words of all senses of the secret word or clue such as synonyms, antonyms, hypernyms, hyponyms, meronyms, holonyms, and verb entailments.

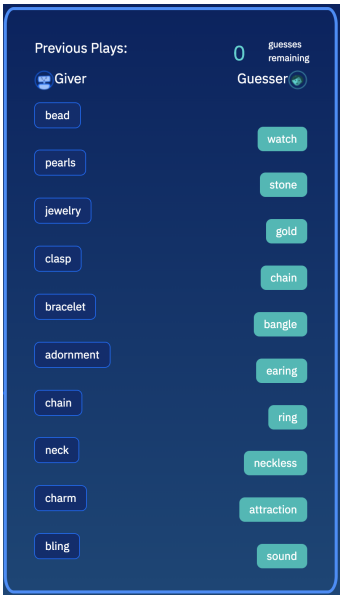
3.1.1 Model A: User Plays as Guesser. Model A generates candidate hints using free association norms, word embeddings, and WordNet [52] (a collection of word level features like antonyms, synonyms, hypernyms etc.) and scores the hints based on a Gradient Boosting Machine (Supervised Machine Learning) model trained on Taboo cards (taboo words as clues). Upon receiving a guess, it reranks the candidates based on which is closer to the target than the previous guess. Upon receiving a hint, the AI finds the intersecting words of the hint and the previous hints based on paths in a knowledge graph and ranks them based on the model. The agent uses a secret word



(a) Model A. User plays as "guesser".

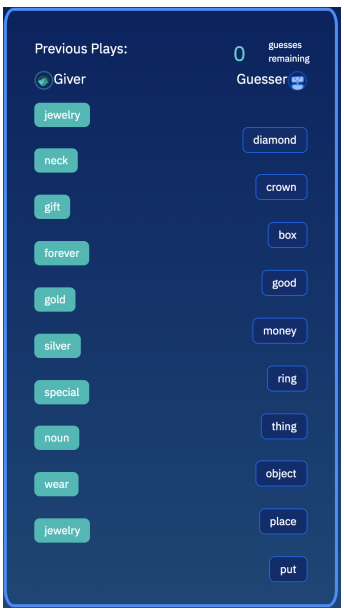


(b) Model B. User plays as "guesser".



(c) Model C. User plays as "guesser".

(d) Model A. User plays as "giver".



(e) Model B. User plays as "giver".

(f) Model C. User plays as "giver".

Figure 1: Examples of gameplay for the six models with the target word "necklace". All users in these examples did not win (since they exhausted their 10 turns). For examples in which user plays as "guesser", guesses appearing on right hand side were given by the user and hints shown on left hand side were generated by the AI agent. For examples in which user plays as "giver", hints appearing on left hand side were given by the user and guesses shown on left hand side were generated by the AI agent.

to generate candidates using the Candidate Generation features (Free Association Norm, WordNet and Word embeddings), scores the candidates based on a GBM model trained on taboo cards with taboo words as hints and outputs the candidate with highest score as next clue. Upon receiving a new guess from the user, the agent re-scores the candidates treating the new guess as secret word and outputs the candidate which is closer to the target than the guess.[1]

[1]The AI response typically takes 1-2 seconds. Across all conditions we added a delay of 2 more additional seconds.

Characteristic	Agent Role	AI Agent	Percentage
Adjusts Hints/Guesses	**Guesser**	Model A	100%
		Model B	100%
		Model C	100%
	Giver	Model A	9%
		Model B	9%
		Model C	0%
Synonym Hints/Guesses	**Guesser**	Model A	21%
		Model B	18%
		Model C	33%
	Giver	Model A	6%
		Model B	6%
		Model C	5%
Antonym Hints/Guesses	**Guesser**	Model A	12%
		Model B	3%
		Model C	12%
	Giver	Model A	2%
		Model B	1%
		Model C	1%

Table 1: AI Agent Characterization of six AI agents in the study. Gameplays sampled include users who exhausted all 10 attempts (i.e. lost the game).

3.1.2 Model A: User plays as "Giver". This agent generates candidates, scores them based on the GBM model trained on Free Association Norm. Upon receiving a clue, the agent finds the intersecting words in the paths of the clue and the previous clues in Conceptnet, scores them based on the GBM model and outputs the candidate with highest score as the next guess. If there are no words intersecting, then it re-scores candidates based on the new clue and outputs the candidate with highest score.

3.2 Model B: Reinforcement Learning

Agent-agent self-play was demonstrated to be effective for agent improvement in many games such as Go [59], Poker [10], Starcraft II [63] and Dota 2 [51]. We convert the GBM models into two end-to-end neural networks which are pre-trained using similar features for inputs and similar training targets as supervised models to equip them with some "common sense", followed by agent-agent self-play as model fine-tuning to try to learn agent's strategies. We "fine-tune" both the pre-trained neural agents with self-play. We use experience replay [39] buffer to store past games and policy gradient [61, 66] for training. Since the game is episodic (we limit the agents to play up to 10 turns per secret word), we are able to select and store the games that are successful and train more on those successful ones and success rates are approximately monotonically increased. Since multi-agent learning suffers from non-stationary [11] issue, we empirically found out that with pre-training in place, the agents could still converge to 92% success rate, with 80% to start with using pre-trained models.

3.2.1 Model B: User plays as "Guesser". The agent is a neural networks policy at turn t as $\pi_{guesser}(a_t|s, g_1, ..., g_{t-1}, c_1, ..., c_{t-1}; \theta)$ modeled as a LSTM [27] with parameters θ. At each LSTM step t, the previous guess g_{t-1}, and the previous clue c_{t-1} are input in the form of feature-concatenated vectors. To pre-train LSTM with Taboo cards, we treat each training example as a 1-step sequence where the previous guess and previous clue are zero-vectors.

3.2.2 Model B: User plays as "Giver". This agent is a neural policy $\pi_{giver}(a_t|g_1, ..., g_{t-1}, c_1, c_t; \phi)$ modeled as another LSTM with parameters ϕ where each step t has g_{t-1} and c_t concatenated as input. Similar to the Model B ("guesser"), this model is pre-trained with trivial "1-step sequence" created from Free Association word pairs. The pairs are sampled according to the FSG (*Forward Association Strength*) scores $P(g|c)$ where c is a FA cue and g is a FA target, and sample c uniformly. We mask the previous guess with zero vectors for pre-training.

3.3 Model C: Data Driven

Instead of creating neural networks that encode word relational knowledge and generalize, we also created two AI agents that are purely data-driven. The relational knowledge we use is the Small World of Words [17] (the pre-processed 2018 data), the most recent and largest collection of word evocation dataset that is created through the word evocation experiment, where a word called "cue" is shown to a participant who is asked to freely come up with another related word called "target". The experiment is usually conducted on many participants for many cue words, and the data produced from the group of people exhibits a collective nature of word relatedness. The probabilities of a target word given a cue word can be obtained from such data and be used as scores for a "giver" or a "guesser".

In evocation data, $P(b|a)$ means when a word a is used as a cue word, how likely it is to propose another word b as a response (target). When shown the word a, one has to go through a thought process to figure out a b. Evocation data is, in essence, an implicit yet "conscious" evaluation of word directional relatedness. Since evocation data is collected from a group of participants [17, 44], each

cue-target pair can be generated from many participants. We could derive a count-based indicator of $P(b|a)$, a score called *Forward Association Strength* (FSG) [44], a well-known metric that is useful to human memory and cognition study. Formally,

$$FSG(b|a) = \frac{count(b \text{ as a response}|a \text{ is cue})}{count(a \text{ is cue})} \quad (1)$$

Thus, we can simply derive "giver" and "guesser" strategies using those FSG scores.

3.3.1 Model C: User plays as "Guesser". The agent uses a score called BSG (backward strength) [44], which is the opposite direction of FSG. Under BSG($b|a$) answers the question of "when a is used as a target word, how likely b is a cue to produce the target". A common practice to obtain BSG is by using FSG (which is mathematically correct when the clue has equal counts, which is the case in SWOW dataset) [44]:

$$BSG(b|a) = \frac{FSG(a|b)}{\sum_c FSG(a|c)} \quad (2)$$

The AI Agent adopts a formula with three variables, a, b, and c. a is a target word, b is a cue word, and c is a range of words, of which b is a member. Thus, the agent is simply using BSG scores, i.e. when given the target word a, the agent is a "greedy" planner that gives clues b in the BSG descending order. Although simple, the strategy guarantees that each clue is maximally informative under the clue-independence assumption. Although we can always design more complex models without that assumption, we observe that such a greedy planner is surprisingly effective in practice (a high win-rate with human) and we left further improvements for future work.

3.3.2 Model C: User plays as "Giver". In turn t in the *Guess the Word*, a set of t clues $\{c_1, ..., c_t\}$ will be provided and under the assumption that each clue has equal importance, a response g_t is chosen among candidate set G by

$$g_t = \arg\max_{g \in G} P(g|c_1, ...c_t) = \arg\max_g \frac{\Pi_{k=1}^t FSG(g|c_k)}{P(g)^{t-1}} \quad (3)$$

The above formulation is derived by applying Bayesian rule twice under one assumption: clues are independent under any condition, i.e. $P(c_1, ..., c_t|X) = \Pi_k^t P(c_k|X)$ where X can be any condition or no condition (marginal distribution). Note that, the denominator is a power of the marginal likelihood $P(g)$, which is the normalized count of a candidate g as a target over the total counts of all targets in the entire SWOW dataset. The intuition of Equation 3 is that the agent tends to choose the g that clues can increase its chance the most for, compared to the prior of the g.

3.4 AI Agent Characterization

The AI agents used to play the *Guess the Word* were developed by a team of researchers. The researchers developed six AI agents, three for the "giver" role (AI has the target word and provides hints to the user) and three for the "guesser" role (user has the target role and provides hints to the AI). While we detailed the technical details of each of the models in the above sections, we also characterize the agents in terms of their behaviors (responsiveness (adjusting hints/guesses), percentage of synonyms used, and percentage of antonyms used). Examples of the different sequences of clues given to the user by the AI for the various models and target words can be seen in Table 3.

Agent	User Role	Score	Number of Turns
Model A	**Guesser**	5.43	3.05
	Giver	5.90	2.21
Model B	**Guesser**	5.55	4.90
	Giver	4.77	3.88
Model C	**Guesser**	6.06	2.37
	Giver	5.47	2.52

Table 2: Score average (out of 10) and Average number of turns (out of 10) for all agents

If we consider the AI agent behavior for one of the models, say the Model ("Guesser"), we can take a systematic approach as done in prior work [21] to characterize the agent's behavior. We can't make the assumption that because the model has access to Small World of Words, it will give the clues in Small World of Words associated with the target word "politician", for example. In Small World of Words, there is rich information about the word politician, i.e. examples of politicians like "Bernie Sanders", characterizations of politicians like "left wing" and other actions that may be associated with a politician like "canvassing", yet the hints that AI agent gives for the Model C for the word politician are: "corrupt", "hypocrite", "representative", "candidate", and "liar".

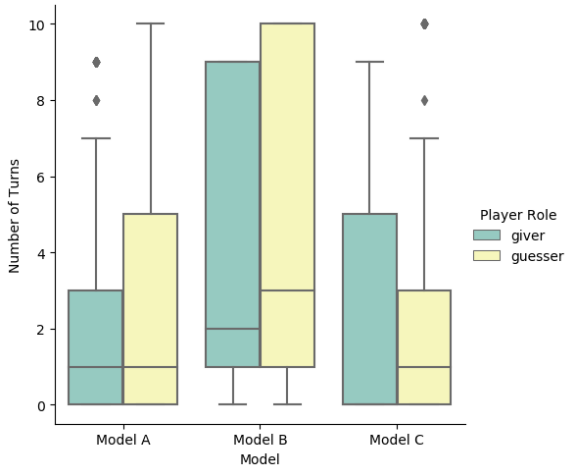


Figure 2: Number of turns plotted for all of the Agents.

To characterize the six variations of *Guess the Word* agents (which we will refer to as Wordbot), we ran a pilot study in which participants were paid a $3.00 amount per 10 games, commensurate to the average time it took them. We analyzed 269 games across 10 target words for the "giver" agents and 312 games across 10 target words for Guesser. These games were part of a pilot study. The sole purpose of the inclusion of these games in this paper is

to characterize the model's behavior beyond equations in a more digestible way to the HCI community. To maximize the amount of guesses/hints per game, we sampled these games because they were not won by the 10th try (they all included a set of 10 guesses or hints depending on the role of the agent).

Adjusting hints. How often do the Wordbots change the hint/guess given the previous guess/hint? We analyzed the guess and hint sequences for each agent. When users play as the "giver", all reactions are unique given that Wordbot is reacting to each individual's hints. This is reflected by the 100% in Table 1. When users played as the "guesser", we calculated the number of sequences that differed from the dominant sequence. For example, for players who exhausted all 10 guesses for the Model B agent, the dominant sequence of guesses for the target word "vanilla" was: 'fruit', 'chocolate', 'flavor', 'bean', 'relish', 'shake', 'plain', 'plant', 'basic', 'standard', 'orchid'. 18 individuals received these hints from the AI, whereas two deviated from the dominant sequence based on the player's guesses. For example one of those sequences looked like this: 'fruit', 'chocolate', 'flavor', 'relish', 'shake', 'plain', 'plant', 'basic', 'standard', 'orchid', 'twist'. We counted the total number of sequences which dominated from the main sequence. 9% of the games adjusted their hints for Model B, 9% of the games had adjusted hints for the Model A, and 0% of the games adjusted their hints for the Model C. When users played as the "guesser", adjustments occur at the very end of the sequence with the guess sequence remaining static for the majority of the guesses (for all agents). Percentages can be seen in Table 1.

Synonym versus antonym hints. Synonyms and antonyms were calculated differently for when users played as "giver" and when users played as the "guesser". When users played as the "guesser", all unique hints provided by the AI agent were marked as either synonym or antonym of the target word. The final percentage presented are the percentage of synonyms or antonyms among the unique hints given. When users played as the "giver", we found a list of synonyms and antonyms of all the hints given by the player to the AI agent and marked whether those synonyms and antonyms appeared in the guesses from the AI agent. The final percentage are the number of those synonyms/antonyms (of the user's hints) that the AI agent supplied as a guess over the entire set of guesses.

4 METHODOLOGY

To answer our research questions, we ran a large-scale on study on Amazon Mechanical Turk. Each participant played 10 rounds and then took a survey. For each round, participants were allowed a maximum of 10 guesses/hints. If the human-AI team was not able to help each other guess the target word correctly after 10 attempts, they lost the round and moved to the next target word. The decision to limit the number of attempts was motivated by findings in previous studies that demonstrate user frustration when the number of attempts are not limited [21]. Based on these previous findings and to reduce user agitation and maximize the use of participant time, we limited the maximum length of the game. We looked at how the following factors impacted user perceptions of their opponent:

- Whether participants perceived their partners to be a AI or a human
- The model with which the participant was playing with (Model A, Model B, Model C)
- The role users were playing ("guesser" vs. "giver")

Participants were either told they were playing with an AI agent or they were playing with another human. Before being assigned to their opponent, they saw a configuration page similar to the configuration page that informed them of the nature of their opponent. Participants assigned to the AI condition were told "Please wait while we load the AI agent you will play against for the next ten rounds". Participants who were assigned the human condition were told "Please wait while we match you with another person for you play the next ten rounds". Upon the beginning the study, participants are told definitively about the AI identity of their partner. Only at the very end of the study are participants asked about the identity and perception of their partners (i.e. intelligence). We measure their impressions after several rounds of gameplay, so that they make an informed overall impression of their partner based on the interactions. We were less interested in the changes of perception while playing the games (though that is an interesting direction for future work) and care more about the fully formed impression after several rounds. Participants were assigned to either play the "guesser" role or the "giver" role. "Guesser" and "giver" agent roles in this study embody various aspects of interactivity. For example, when an individual plays as the "giver" agent, the AI agent is more responsive since it responds to the user's clues and adjusts its guesses based on the clues a user gives. The user has more control and agency since it is able to "steer" the AI since the AI adjusts its behavior. When the user plays as the "guesser", the AI does very little adjusting to steer the individual in the right direction and continues to give clues based on its personal knowledge base. Based on these characteristics, it is fair to hypothesize that users will find the agents more likeable when they play the role of giver. We consider these differences as we investigate the social perceptions of the AI agents in this study. Participants were also assigned to a particular model. They were either assigned to the Model A, Model B, or Model C.

For all players, we used one word list of ten words and balanced it for difficulty: "necklace","strong", "hair", "book", "dog", "vanilla", "baby", "cold", "politician", and "house". Similar prior work [21] used a similar metric (accessibility index of words, a measure from [44]) to balance for word difficulty. Gero et. al compared user mental models of AI agents in a collaborative word game to win rate [21]. Their study did not show differences between different words lists which were balanced for word difficulty. The game was developed into an online web application using Flask (a lightweight Python framework for web apps) and React (a Javascript library for building front-end interfaces). In pilot studies, the average time of completion was 25 minutes. Based on this all participants were paid $3.50 commensurate with federal minimum wage.

5 SURVEY INSTRUMENT

5.1 Dependent Variables: User Perception of Opponent

To address our research questions, we assessed user perception of rapport, likeability, and intelligence of the opponent. Based on previous work [4, 46], we asked participants to indicate how much

Target Word	AI Agent	Clues
Strong		
	Model A	weak, strength, powerful, firm, forceful, might, courage, hardy, superman, power, overpower
	Model B	hard, need, loud, love, tough, firm, bitter, steel, fit, powerful, good
	Model C	powerful, potent, mighty, resilient, muscular, weak, robust, sturdy, forceful, durable, manly
Politician		
	Model A	corrupt, china, apathetic, campaign, launch, vice, operation, governor, rescue, whom, strategy
	Model B	elect, official, government, democrat, congress, change, president, person, honest, republican, mayor
	Model C	corrupt, hypocrite, representative, candidate, liar, senator, diplomat, untrustworthy, elect, governor, prominent
Vanilla		
	Model A	chocolate, sparrow, conversation, dark, tapioca, tangy, lift, forbidden, advice, pirate, gloomy
	Model B	bark, fruit, chocolate, flavor, bean, relish, shake, plain, plant, basic, standard
	Model C	extract, milkshake, pudding, flavor, plain, custard, yoghurt, essence, tapioca, imitation, bourbon
Cold		
	Model A	snow, frozen, virus, freeze, hot, winter, sneeze, warm, frigid, frost, chill
	Model B	chill, blood, snow, frigid, frost, hot, sick, winter, fever, ice, temperature
	Model C	freezing, shiver, chilly, frigid, antarctic, fridge, freezer, icy, refrigerator, alaska, ice

Table 3: Examples of different hints given by the AI when the user plays as "guesser" for the different models.

they agreed/disagreed with statements like, "My opponent was not paying attention to me.", "My opponent and I worked towards a common goal.", and "I feel that my opponent trusts me.". To measure the other dimensions in our research questions (likeability, intelligence, rapport), we used a list of semantic differential scales. We adapted scales on these dimensions by [8, 47, 56]. Participants were asked to rate their opponent on pairs of antonyms (i.e. unfriendly/friendly, unpleasant/likeable, ignorant/knowledgable). All of the perception questions were asked based on a 7-point Likert scale. The averages for perception dimensions were calculated for analysis. Prior work has shown an appropriate Cronbach alpha for these measures [4]. Below, we list the dependent variables measured in the post-survey.

- **Intelligence** To measure intelligence, we used a list of four semantic differential scales also used in [8, 47, 56], in which participants rated their opponent on a team 7 point Likert scale as Unintelligent/Intelligent, Ignorant/Knowledgeable, Incompetent/Competent, and Irresponsible/Responsible. The intelligence value was an average of these four scales.
- **Rapport** We measured rapport by adapting an instrument from [46], in which participants rated items like "My opponent seemed engaged" or "My opponent and I worked towards a common goal" on a 7 point Likert scale. To measure rapport, we asked nine questions in which participants responded with Strongly Disagree/Strongly Agree.
- **Likeability** To measure likeability, we used five semantic differential scales, also used in [8, 47, 56] in which participants rated their opponent on a 7 point Likert scale as unfriendly/friendly, not kind/kind, unpleasant/pleasant, not cheerful/cheerful and dissimilar to me/similar to me.

5.2 AI Score: Perceived Partner Type

To ascertain whether participants felt that they were competing against a human or an AI, we collected an AI score as done in [3, 30], in which we asked participants about whether they believed they were interacting with a human or an AI. The AI score is calculated based on the average of two questions in our post survey that asked participants about whether they perceived their opponent to be human or AI (on a 7 point Likert scale). We then calculated the median of the score for all participants (Mdn=4) in the study and segmented users based on their scores into Human or AI. Prior work shows that in studies in which users are told they are interacting with humans/AI and they are not (i.e. some level of deception involved), participants do not necessarily believe the conditions [3, 57]. We were interested in the perceived AI score, as prior work has already investigated the suspicion effect [30] that occurs when people are suspicious when we tell them they are interacting with a human, or when we do not disclose with whom they are interacting. For this reason, our analysis included perceived AI identity (human, AI) based on the AI score collected only.

Agent	User Role	Assigned AI Identity	N
Model A	**Guesser**	AI	12
		Human	16
	Giver	AI	13
		Human	14
Model B	**Guesser**	AI	12
		Human	24
	Giver	AI	38
		Human	15
Model C	**Guesser**	AI	8
		Human	20
	Giver	AI	19
		Human	8

Table 4: Number of participants per treatment.

6 RESULTS

6.1 Participants

A total of 237 participants participated in the study. We performed several attentiveness tests to preserve the integrity of the data collected. We excluded those who did not pass the linguistic attentiveness task [42]. This left us with 199 subjects. The demographic variables collected: education, age and language can be seen in Table 5. Participants were randomly assigned to the conditions leaving us with the breakdowns seen in Table 4.

Demographic	
Age	26-35 (52%), 36-45 (24%), 45+ (16%), 18-25 (8%)
Language	English (94%), Other (not specified) (3%), Portuguese (1%), Chinese (1%), German (1%)
Education	Middle School (1%), High School (28%), Bachelors (61%), Advanced (10%)

Table 5: Participant Demographics, N=237

6.2 Subjective Social Perception of Opponent

When people perceive their partner to be an AI or human in varying contexts, do they perceive one as more intelligent, likeable, and have more rapport with one than the other? We conducted a three way ANOVA to compare the effects of perceived AI identity (human vs. AI), user role ("giver" vs. "guesser"), and model (Model A vs. Model B vs. Model C) and their interaction on perceived intelligence, rapport, and likeability. Prior work shows that only a subset of individuals believed the deceptive condition (telling users AI agent is human) [57]. Similarly, in our study, only 45% of participants believed that the bot was an AI when we told users it was an AI. For this reason, we included an AI score to answer our research questions around perceived AI identity. The details of the each ANOVA are included in Table 6, with details in the subsections below. For all post-hoc analyses, statistical significance was accepted at the $p < 0.025$ level for simple two-way interactions and simple main effects.

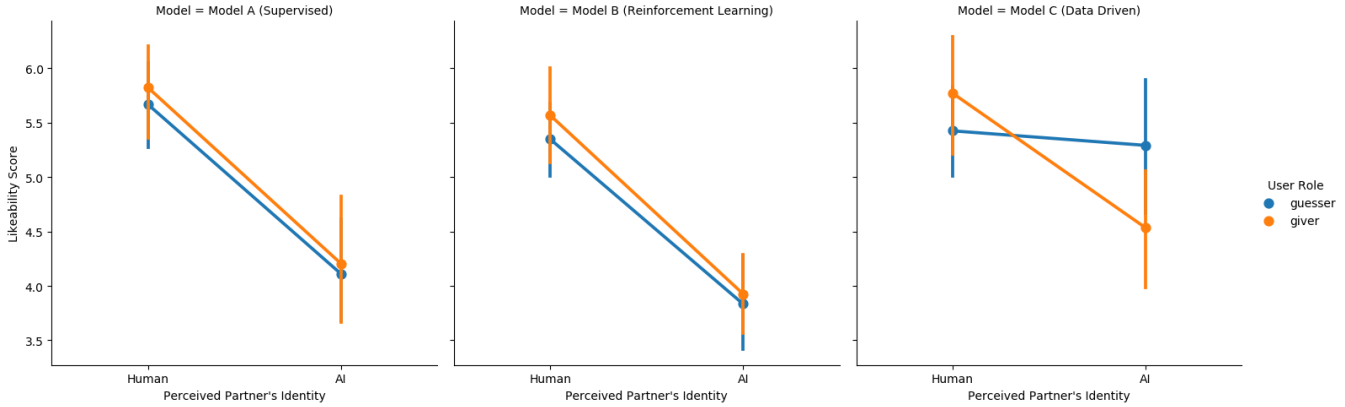
6.2.1 Intelligence. A three-way ANOVA was conducted to determine the effects of perceived AI identity, model and player role on perceived intelligence. There was a statistically significant three-way interaction between perceived AI identity, player role, and model $F(11,187) = 6.96$, $p < 0.001$. There was a statistically significant simple two-way interaction between perceived AI identity and player role for Model B, $F(3, 85) = 10.0$, $p < 0.001$, and for the Model A $F(3,51) = 6.53$, $p < 0.001$, but not for the Model C, $F(3, 51) = 2.68$, $p = 0.06$. For the Model B and Model A, this result suggests that the effect of the user role ("giver" vs. "guesser") on perceived intelligence depends on the perceived AI identity of the agent.

There was a statistically significant simple main effect of user role ("giver" vs. "guesser") for the Model A when individuals perceived their partner to be an AI. In other words, the mean perceived intelligence score for "giver" and "guesser" roles was statistically significant for users who interacted with the Model A and perceived their partners to be an AI. All simple pairwise comparisons, between the different roles, were run for individuals who interacted with the Model A and perceived their partners to be an AI agent with a Bonferroni adjustment applied. For those interacting with the Model A and perceiving their partners to be AI, there was a statistically significant mean difference between perceived intelligence of partners for those who played as "guesser" (4.46 ± 1.82) and those who played as "giver" (5.56 ± 0.95) of 1.1 ($p < 0.05$) (seen in Figure 3a). For those who believed their partners to be an AI, users playing Model A ("giver") found their partners to be more intelligent than those playing Model A ("guesser").

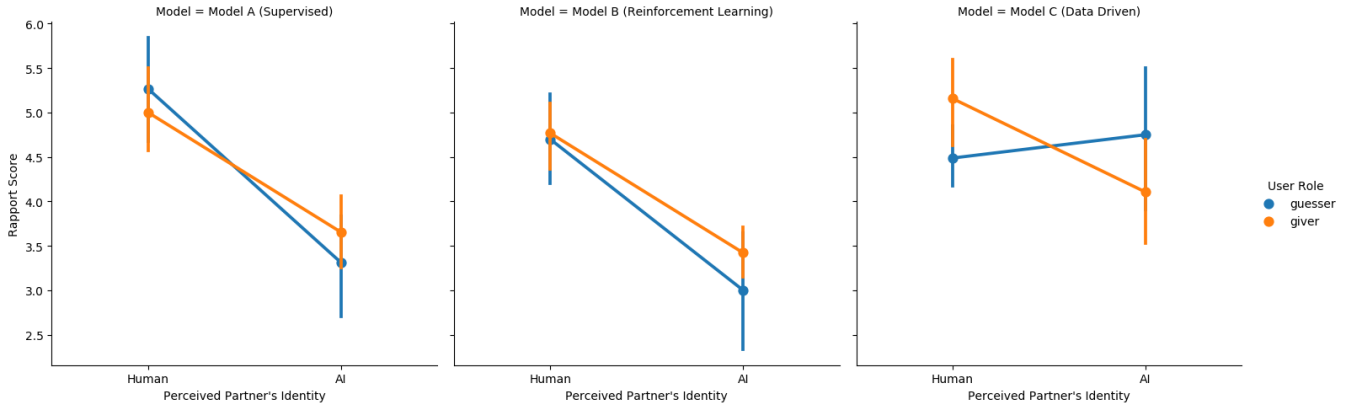
6.2.2 Likeability. As with intelligence (and all other subjective social perception measures), a three-way ANOVA was conducted to determine the effects of perceived AI identity, model and player role on perceived likeability. There was a statistically significant three-way interaction between perceived AI identity, player role, and model, $F(11, 187) = 10.33$, $p < 0.001$. There was a statistically significant simple two-way interaction between the model and the player role when the agent was perceived to be an AI, $F(5, 96) = 2.80$, $p = 0.02$, but not when the partner was perceived to be a human, $F(5, 91) = 0.75$, $p = 0.588$. When the partner is perceived to be an AI, this result suggests that the effect of the player role (giver vs. guesser) on perceived likeability depends on the model with which it is interacting. There was a statistically significant simple main effect of model for individuals who believed they were playing with an AI agent as a guesser, $F(2, 187) = 5.5$, $p = 0.004$. For those interacting as "guesser" and perceiving their partners to be an AI agent, there was a statistically significant mean difference of perceived likeability of partners between those who played with the Model C agent (5.29 ± 0.95) and those who played with the Model B agent (3.83 ± 0.80) of 1.46 ($p< 0.01$) and between the Model C agent (5.29 ± 0.95) and the Model A agent (4.11 ± 0.89) of 1.18 ($p<0.05$) (seen in Figure Figure 3b). Simply put, we observe a drop in perceived likeability for both the Model B ("guesser") and the Model A ("guesser") agent compared to the Model C ("guesser") when it is perceived as an AI.

.

(a) Perceived intelligence of partners measured after cooperative gameplay. Intelligence score is based on the mean of 4 items from the survey (1=Not Intelligent, 7=Intelligent). For those interacting with the Model A and perceiving their partners to be AI, there was a statistically significant mean difference between perceived intelligence of partners for those who played as "guesser" (4.46 ± 1.82) and those who played as "giver" (5.56 ± 0.95) ($p < 0.05$), with those playing as the "giver" perceiving their partners to be more intelligent.



(b) Perceived likeability of partners measured after cooperative gameplay. Likeability score is based on the mean of 6 items from the survey (1=Not Likeable, 7=Likeable). For those interacting as "guesser" and perceiving their partners to be an AI agent, there was a statistically significant mean difference of perceived likeability of partners between those who played with the Model C agent (5.29 ± 0.95) and those who played with the Model B agent (3.83 ± 0.80) of 1.46 ($p< 0.01$) and between the Model C agent (5.29 ± 0.95) and the Model A agent (4.11 ± 0.89) of 1.18 ($p<0.05$).



(c) Perceived rapport of partners measured after cooperative gameplay. Rapport score is based on the mean of 9 items from the survey (1=No Rapport, 7=Rapport). For those interacting as "guesser" and perceiving their partners to be AI, there was a statistically significant mean difference of perceived rapport of partners between those who played with the Model C agent (4.75 ± 1.26) and those who played with the Model B agent (3.00 ± 1.25) of 1.75 ($p< 0.001$) and between the Model C agent (4.75 ± 1.26) and the Model A agent (3.31 ± 1.03) ($p<0.01$).

Figure 3: The social perception measures (intelligence, rapport, and likeability) plotted against the perceived roles of the agent. Different hues represent the different roles users were assigned to play against the agent.

Dependent Variable	Source	SS	df	F
Intelligence	User Role x Model x Perceived Role	91.86	11.0	6.96***
Likeability	User Role x Model x Perceived Role	111.20	11.0	10.33***
Rapport	User Role x Model x Perceived Role	103.08	11.0	8.57***
Gameplay	User Role x Model x Perceived Role	235.97	11.0	7.85***

Significance Codes : ***p <0.001, **p <0.01, *p <0.05

Table 6: ANOVA predicting dependent variables (intelligence, likeability, rapport, creativity) based on assigned conditions: ("guesser" vs. "giver"), model (Model A, Model B, Model C), and perceived role (AI, Human).

6.2.3 Rapport. A three-way ANOVA was conducted to determine the effects of perceived AI identity, model and player role on perceived rapport. There was a statistically significant three-way interaction between perceived AI identity, player role, and model, $F(11, 187) = 8.57$, $p < 0.001$.

There was a statistically significant simple two-way interaction between the model and the player role when the agent was perceived to be an AI, $F(5, 96) = 3.89$, $p = 0.003$, but not when the partner was perceived to be a human, $F(5, 91) = 1.34$, $p = 0.25$. When the partner is perceived to be an AI, this result suggests that the effect of the player role ("giver" vs. "guesser") on perceived rapport depends on the model with which it is interacting. There was a statistically significant simple main effect of model for individuals who believed they were playing with an AI agent as a guesser, $F(2, 187) = 7.23$, $p < 0.001$. For those interacting as "guesser" and perceiving their partners to be AI, there was a statistically significant mean difference of perceived rapport of partners between those who played with the Model C agent (4.75 ± 1.26) and those who played with the Model B agent (3.00 ± 1.25) of 1.75 ($p< 0.001$) and between the Model C agent (4.75 ± 1.26) and the Model A agent (3.31 ± 1.03) of 1.44 ($p<0.01$). Simply put, we observe a drop in perceived rapport for both the Model B "guesser" and the Model A "guesser" agent when it is perceived as an AI when compared to the Model C "guesser".

6.3 Game Play Results

A three-way ANOVA was conducted to determine the effects of perceived AI identity, model and player role on gameplay outcome. We operationalized gameplay outcome as average number of turns during gameplay (i.e. the higher the number of turns users took, the worse the user performance). These gameplay outcome for the different models can be seen in Table 2 and Figure 2. There was a statistically significant three-way interaction between perceived AI identity, player role, and model, $F(11, 187) = 7.85$, $p < 0.001$.

There was a statistically significant simple two-way interaction between the player role ("giver" vs. "guesser") and the perceived AI identity (AI vs. Human) for Model B $F(3,85) = 6.18$ $p< 0.001$ and for the Model A $F(5,51) = 3.26$ $p < 0.05$, but not for the Model C $F(3,51) = 0.992$ $p = 0.404$. When users interacted with the Model A or Model B, this result suggests that the effect of the player role ("giver" vs. "guesser") on number of turns taken during the game depends on the perceived AI identity.

There was a statistically significant simple main effect of perceived AI identity for individuals who played against Model B ("guesser") $F(1,187) = 6.48$ $p < 0.05$ and Model A ("giver") $F(1,187) = 5.92$ $p < 0.05$. In other words, for those interacting as "guesser" and playing with Model B, there was a statistically significant mean difference of number of turns between those who perceived their partners to be an AI (3.91 ± 0.92)and those who perceived their partners to be a Human (5.40 ± 2.15) ($p< 0.05$).

For those interacting as "giver" and playing with the Model A, there was a statistically significant mean difference of number of turns between those who perceived their partners to be an AI (1.41 ± 0.35) and those who perceived their partners to be a Human (2.96 ± 2.27) ($p< 0.05$), with those believing that their partners to be an AI taking fewer turns to win than those who believed to be interacting with human partner. Differences can be seen in Figure 4.

6.4 Potential Underlying Reasons for Differences

Given the results from our analysis, particularly around the effect of perceived identity on the overall game performance, we wanted to further investigate why we were seeing these differences. Since we saw statistically significant differences for perceived identity for those interacting with the Model A("Giver") and Model A("Guesser"), we looked at top 20% games (yielding 60 games from 6 users) from high performing users who believed they were playing with the AI and bottom 20% of games (yielding 60 games from 6 users) from low performing users who believed they were playing with a human. We wanted to identify different ways people were interacting with their partners and differences in the kinds of hints people were using. Two of the authors individually looked at each set of 60 games and identified a codebook of trends in the kinds of words used. The authors compared trends and finalized a list. We found that those who believed their partner to be a human repeated clues, used relational words like "opposite" to indicate a word was the antonym of the target word and used suggestive hints to interact with their partners. All of these tactics contribute to an increase in number of turns taken to win that can impact the overall performance of users. The list of tactics used is presented in Table 7.

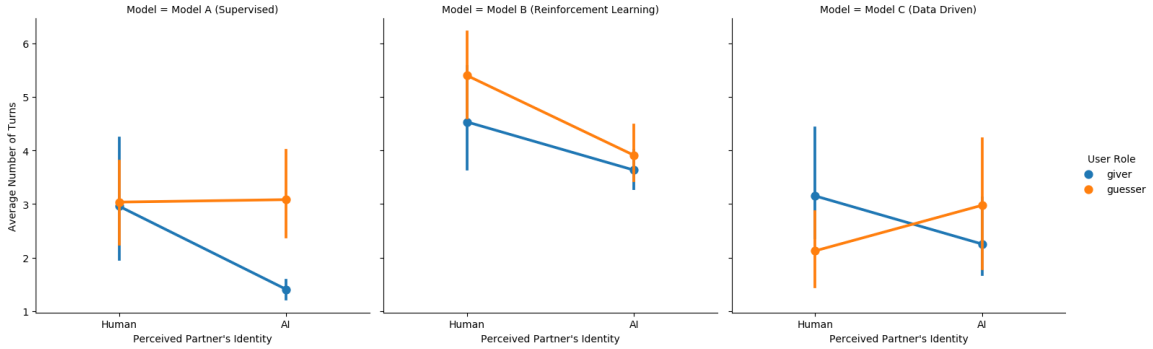


Figure 4: Average number of turns plotted against the perceived roles of the agent. Different hues represent the different roles users were assigned to play against the agent. For those interacting as "guesser" and playing with Model B, there was a statistically significant mean difference of number of turns between those who perceived their partners to be an AI (3.91 ± 0.92) and those who perceived their partners to be a Human (5.40 ± 2.15) ($p < 0.05$). For those interacting as "giver" and playing with the Model A, there was a statistically significant mean difference of number of turns between those who perceived their partners to be an AI (1.41 ± 0.35) and those who perceived their partners to be a Human (2.96 ± 2.27) ($p < 0.05$).

Perceived Partner	Trend	Definition	Example Target: Hints
AI	Trigger words	Trigger words are words that trigger the target word in one turn	**necklace**: pearl; **baby**: infant
	Diverse words	Using adjectives to describe target word, synonyms, antonyms (a variety of different kinds of descriptive words)	**vanilla**: chocolate, bean, strawberry, white, cream, flavor, taste, syrup
Human	Repetition	Individuals repeat the same words	**necklace**:jewel, girl, gold, neck, jewel, neck, chain, neck, chain, jewel
	Relational words	Using multiple turns to and relating one word to another, i.e. "opposite" to indicate the hint sent prior is the "opposite" of the target word	**cold**: hot, heat, opposite, hot
	Suggestive Hints	Individuals trying to learn their partner's gender and using suggestive terms unrelated to the target word	**vanilla**: ice, snow, hot, sure, cream, opposite, human, male, female, sex

Table 7: Types of behavior users exhibited when they believed to be interacting with a human versus when they believed to be interacting with an AI and Human for Model A when player interacted as "giver"

6.5 Limitations

Our study has a few limitations. First, as many researchers have done in the past, we use an AI-driven cooperative game with partially missing information as a testbed for our evaluation [4, 21]. To consider the ecologicial validity of our results, and thus generalizeability we must consider the environment in which we ran these studies. Our findings hold credence given that the game in question is a cooperative one with missing information (that resembles human-AI interactive scenarios) and not a zero-sum game. Given the environment in which we ran our study, there are potential limitations to the generalizability of our results to other human-AI environments, but certainly of value to other human-AI collaborative spaces in which the user may perceive an AI partner to be a human or an AI. Secondly, we use Mechanical Turkers who are motivated to finish the study as fast as they possibly can. To account for this, we filtered out users who did not appear to put in meaningful effort in gameplay (filler words as guesses). Additionally, the top 20% of games from high performing users in the AI group and the bottom 20% of games from low performing users in the human group were compared. While we acknowledge this approach has limitations since all interactions were not analyzed, the results of this analysis show the most extreme cases in each of these groups and help readers better understand how users interacted at a granular level.

7 DISCUSSION

Our results demonstrate that there are social perception differences between different AI agents, as there are when the direction of communication is varied. The direction of the communication with the AI agent combined with perception of AI identity impacts both gameplay performance and social perception. We discuss our findings below.

7.1 Explaining Differences in Performance

We observed that when people think they are interacting with a human, they employ different strategies to communicate the importance of words in relation to the target word and the relationship of a word to the target word. Our analyses revealed that for two of the models (Model B (guesser) and Model A (giver)), there were statistically significant differences for the average number of turns taken by participants who perceive their partners to be human compared to those who perceive their partners to be an AI. Looking at the types of words players use to communicate with their partners, we observe that users are trying to leverage the one word communication medium to communicate richer information that is not interpretable by an AI, i.e. "opposite" to imply that that the prior word sent was an antonym of the target word, or repeating words to highlight the importance of words. Here, we see an overestimation of their partner's ability to interpret. Prior work has shown that an individual who overestimates their AI partner's abilities performs worse in such collaborative spaces [21]. Belief that the partner is a human is a form of overestimation. By overestimating their partner's abilities, user's gameplay performance deteriorated because they attempted to communicate information that was not interpretable by the AI agent.

7.2 Interpreting Social Perception Differences

We observe many different relationship trends between models, perceived AI identity, and direction of communication. Let's consider the measure of intelligence, for example. We find that there are statistically significant differences between perceived intelligence score between "giver" and "guesser" agents when the AI is perceived to be an AI for the Model A, with the "giver" being perceived as more intelligent than the "guesser" model. We discuss our findings in detail below.

7.2.1 Social Perception Differences for "Giver" and "Guesser": User Control and Direction of Communication. In the AI-driven cooperative game with partially observable information presented in this paper, *Guess the Word*, there are two roles played: one in which the human must interpret the limited information being communicated by the AI and submit guesses accordingly ("guesser"), and one in which the human sends clues that are digestible and interpretable to an AI agent to be interpreted ("giver"). Our findings show that for one of the AI agents (Model A), when participants believed to be playing against an AI, they found their partners to be more intelligent when they were playing the role of "Giver" then when they were playing the role of "Guesser". Prior work has shown there to be a bias against bots when they disclose their identity [29, 57], but in our experiment, we observe the bias against the AI to be more present when users play the "guesser" versus the "giver".

When a player interacts as the "giver", the AI is reactionary, i.e. responding to the user at every turn. Conversely, when a player interacts as the "guesser", they are following the AI's lead, reacting to the the AI's clues. One potential explanation for the differences in social perception could be explained around a user's feeling like they are in control when they play as the "giver" and not when they play as the "guesser". Prior work has identified three components of interactivity that may impact social perceptions: direction of communication, user control, and time [38]. Direction of the communication varies in "giver" and "guesser" roles as does the amount of control a user has to influence the AI's responses. Users have more control as givers since the AI reacts to their clues.

7.2.2 Social Perception Differences for the Different Models. When users play the role of "guesser" and believe their partners to be an AI, the Model C agent is perceived more positively (likeability and rapport) than Model A and Model B. Of all three agents, Model C ("guesser") has a higher average score and lower average number of turns than both the Model A ("guesser") and Model B ("guesser"). One potential explanation could be that it is a better model than the others, one that is better at giving clues that are interpretable by users. The average score (number of turns) for the Model C ("guesser") is higher than all of the others (M=6.06). The average number of turns taken to win are also lower than all of the others (M=2.37). Our results seem to suggest that if the human-AI collaboration is successful (i.e. fewer number of turns, higher average score) that users judge the AI less harshly than they would if they felt that the AI was underperforming, that any "mistake" on the AI's part would result in a more negative perception of the AI, whereas people are more forgiving of their percieved human partners. This is in opposition to prior studies that have discovered a bias against AI and suggests it is not a general bias but a bias against specific AI behavior that can be mitigated through agent-tuning behavior (as shown across 3 different AI agents here).

Our results highlight that it is not enough to evaluate social perceptions with one AI model, as the results across our three AI agents vary based on the behavior of the AI agent. Many researchers in the CHI community investigating human-AI interaction have used one model to measure user perception in conversational agents [43] or other computer-mediated communications that are AI driven [30]. And while those in AI communities often compare models, this is typically in the context of accuracy, performance, etc., and very often in simulation environments. However, our results show that social perceptions vary given the behavior of the AI and the context. To draw conclusions about the user perceptions of the AI agent in a human-AI collaborative environment, it is important to note the context of the collaboration and interactivity, the direction of the communication, and the underlying model driving the interaction. We see very different trends for our Model C when compared to both the Model A and Model B, particularly for the "guesser" role. These differences are explained by how the AI behaves differently and the quality of the clues it provides for users. Whereas with the other "guesser" AI agents we see a decrease in social perception (likeability, rapport, intelligence) when users believe they are interacting with an AI (vs. a Human), we do not observe this trend with the Model C "guesser" AI, suggesting there is something about the the Model C "guesser" AI that impacts how users perceive it.

7.3 Directionality of Communication and Implications for Practitioners

Our findings show that in human-AI interaction, the directionality of communication has an impact on the perception of an agent. Often the perception of the agent in an interaction has an impact on the outcome of the interaction. For this reason, it is important to consider the implications of directionality of communication and the potential implications for practitioners. In human-AI collaborative settings (conversational agents [9], pedagogical systems [23], or co-creation tools [31, 32]), AI systems can be trained to react to human input rather than lead the collaboration and interaction. Further research is required to investigate directionality of communication in settings beyond the *Guess the Word* environment we have investigated in this study, but our findings imply that directionality has an impact on social perception of the AI.

8 CONCLUSION

In this work, we find performance differences given user's perceived identity of their AI partner and the direction of communication. Through an online study in which participants played an AI-driven cooperative game with partially observable information with multiple AI agents in various directions of communication (as "guesser" and "giver"), we show that there are social perception differences with different AI agents, as there are when the direction of communication is varied. We also find that the bias against the AI that has been demonstrated in prior studies [4, 20] varies with the direction of communication and with the AI agent. This research leads to new insights about how to study human-AI collaboration and lays the groundwork for future studies.

REFERENCES

[1] Kemo Adrian, Aysenur Bilgin, Paul Van Eecke, and others. 2016. A Semantic Distance based Architecture for a Guesser Agent in ESSENCE's Location Taboo Challenge. *DIVERSITY@ ECAI* (2016), 33–39.

[2] Ines Arous, Jie Yang, Mourad Khayati, and Philippe Cudré-Mauroux. 2020. OpenCrowd: A Human-AI Collaborative Approach for Finding Social Influencers via Open-Ended Answers Aggregation. In *Proceedings of The Web Conference 2020*. 1851–1862.

[3] Zahra Ashktorab, Q Vera Liao, Casey Dugan, James Johnson, Qian Pan, Wei Zhang, Sadhana Kumaravel, and Murray Campbell. 2020a. Human-ai collaboration in a cooperative game setting: Measuring social perception and outcomes. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–20.

[4] Zahra Ashktorab, Vera Q Liao, Casey Dugan, James Johnson, Qian Pan, Wei Zhang, Sadhana Kumaravel, and Murray Campbell. 2020b. Human-AI Collaboration in a Cooperative Game Setting: Measuring Social Perception and Outcomes. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 96.

[5] Gagan Bansal, Besmira Nushi, Ece Kamar, Daniel S Weld, Walter S Lasecki, and Eric Horvitz. 2019. Updates in human-ai teams: Understanding and addressing the performance/compatibility tradeoff. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 2429–2437.

[6] Nolan Bard, Jakob N Foerster, Sarath Chandar, Neil Burch, Marc Lanctot, H Francis Song, Emilio Parisotto, Vincent Dumoulin, Subhodeep Moitra, Edward Hughes, and others. 2019. The Hanabi Challenge: A New Frontier for AI Research. *arXiv preprint arXiv:1902.00506* (2019).

[7] Nolan Bard, Jakob N Foerster, Sarath Chandar, Neil Burch, Marc Lanctot, H Francis Song, Emilio Parisotto, Vincent Dumoulin, Subhodeep Moitra, Edward Hughes, and others. 2020. The hanabi challenge: A new frontier for ai research. *Artificial Intelligence* 280 (2020), 103216.

[8] Christoph Bartneck, Dana Kulić, Elizabeth Croft, and Susana Zoghbi. 2009. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International journal of social robotics* 1, 1 (2009), 71–81.

[9] Timothy W Bickmore, Laura M Pfeifer, and Brian W Jack. 2009. Taking the time to care: empowering low health literacy hospital patients with virtual nurse agents. In *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 1265–1274.

[10] Noam Brown and Tuomas Sandholm. 2018. Superhuman AI for heads-up no-limit poker: Libratus beats top professionals. *Science* 359, 6374 (2018), 418–424.

[11] Lucian Bu, Robert Babu, Bart De Schutter, and others. 2008. A comprehensive survey of multiagent reinforcement learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 38, 2 (2008), 156–172.

[12] Carrie J Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. 2019. " Hello AI": Uncovering the Onboarding Needs of Medical Practitioners for Human-AI Collaborative Decision-Making. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–24.

[13] Murray Campbell, A Joseph Hoane Jr, and Feng-hsiung Hsu. 2002. Deep blue. *Artificial intelligence* 134, 1-2 (2002), 57–83.

[14] James A Crowder, John Carbone, and Shelli Friess. 2020. Human–AI Collaboration. In *Artificial Psychology*. Springer, 35–50.

[15] Abhishek Das, Satwik Kottur, José MF Moura, Stefan Lee, and Dhruv Batra. 2017a. Learning cooperative visual dialog agents with deep reinforcement learning. In *Proceedings of the IEEE International Conference on Computer Vision*. 2951–2960.

[16] Abhishek Das, Satwik Kottur, Jose M. F. Moura, Stefan Lee, and Dhruv Batra. 2017b. Learning Cooperative Visual Dialog Agents With Deep Reinforcement Learning. In *The IEEE International Conference on Computer Vision (ICCV)*.

[17] Simon De Deyne, Danielle J Navarro, Amy Perfors, Marc Brysbaert, and Gert Storms. 2019. The "Small World of Words" English word association norms for over 12,000 cue words. *Behavior research methods* 51, 3 (2019), 987–1006.

[18] Harm De Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron Courville. 2017. Guesswhat?! visual object discovery through multimodal dialogue. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5503–5512.

[19] Edward J Downes and Sally J McMillan. 2000. Defining interactivity: A qualitative identification of key dimensions. *New media & society* 2, 2 (2000), 157–179.

[20] Ethan Fast and Eric Horvitz. 2016. Long-term trends in the public perception of artificial intelligence. *arXiv preprint arXiv:1609.04904* (2016).

[21] Katy Gero, Zahra Ashktorab, Casey Dugan, and Werner Geyer. Mental Models of AI Agents in a Cooperative Game Setting. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM.

[22] Elizabeth Gibney. 2016. Google AI algorithm masters ancient game of Go. *Nature News* 529, 7587 (2016), 445.

[23] Arthur C Graesser, Patrick Chipman, Brian C Haynes, and Andrew Olney. 2005. AutoTutor: An intelligent tutoring system with mixed-initiative dialogue. *IEEE Transactions on Education* 48, 4 (2005), 612–618.

[24] Francisco J Gutierrez, Sergio F Ochoa, Raymundo Cornejo, and Julita Vassileva. 2019. Designing Computer-Supported Technology to Mediate Intergenerational Social Interaction: A Cultural Perspective. In *Perspectives on Human-Computer Interaction Research with Older People*. Springer, 199–214.

[25] Serhii Havrylov and Ivan Titov. 2017. Emergence of language with multi-agent games: Learning to communicate with sequences of symbols. In *Advances in neural information processing systems*. 2149–2159.

[26] He He, Anusha Balakrishnan, Mihail Eric, and Percy Liang. 2017. Learning symmetric collaborative dialogue agents with dynamic knowledge graph embeddings. *arXiv preprint arXiv:1704.07130* (2017).

[27] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[28] Huang Hu, Xianchao Wu, Bingfeng Luo, Chongyang Tao, Can Xu, Wei Wu, and Zhan Chen. 2018. Playing 20 Question Game with Policy-Based Reinforcement Learning. *arXiv preprint arXiv:1808.07645* (2018).

[29] Fatimah Ishowo-Oloko, Jean-François Bonnefon, Zakariyah Soroye, Jacob Crandall, Iyad Rahwan, and Talal Rahwan. 2019. Behavioural evidence for a transparency–efficiency tradeoff in human–machine cooperation. *Nature Machine Intelligence* 1, 11 (2019), 517–521.

[30] Maurice Jakesch, Megan French, Xiao Ma, Jeffrey T Hancock, and Mor Naaman. 2019. AI-mediated communication: How the perception that profile text was written by AI affects trustworthiness. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–13.

[31] Anna Kantosalo and Hannu Toivonen. 2016. Modes for creative human-computer collaboration: Alternating and task-divided co-creativity. In *Proceedings of the seventh international conference on computational creativity*. 77–84.

[32] Pegah Karimi, Jeba Rezwana, Safat Siddiqui, Mary Lou Maher, and Nasrin Dehbozorgi. 2020. Creative sketching partner: an analysis of human-AI co-creativity. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*. 221–230.

[33] Marc Lanctot, Vinicius Zambaldi, Audrunas Gruslys, Angeliki Lazaridou, Karl Tuyls, Julien Pérolat, David Silver, and Thore Graepel. 2017. A unified game-theoretic approach to multiagent reinforcement learning. In *Advances in Neural Information Processing Systems*. 4190–4203.

[34] Angeliki Lazaridou, Alexander Peysakhovich, and Marco Baroni. 2017. Multi-agent cooperation and the emergence of (natural) language. (2017).

[35] Adam Lerer, Hengyuan Hu, Jakob Foerster, and Noam Brown. Search in Cooperative Partially-Observable Games. (????).

[36] Jiwei Li, Will Monroe, Alan Ritter, Dan Jurafsky, Michel Galley, and Jianfeng Gao. 2016. Deep Reinforcement Learning for Dialogue Generation. In *Proceedings*

of the 2016 Conference on Empirical Methods in Natural Language Processing. 1192–1202.

[37] Claire Liang, Julia Proft, Erik Andersen, and Ross A Knepper. 2019. Implicit communication of actionable information in human-ai teams. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–13.

[38] Sally J McMillan and Jang-Sun Hwang. 2002. Measures of perceived interactivity: An exploration of the role of direction of communication, user control, and time in shaping perceptions of interactivity. *Journal of advertising* 31, 3 (2002), 29–42.

[39] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. 2013. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602* (2013).

[40] Yi Mou and Kun Xu. 2017. The media inequality: Comparing the initial human-human and human-AI social interactions. *Computers in Human Behavior* 72 (2017), 432–440.

[41] Michael Muller, Susan R Fussell, Ge Gao, Pamela J Hinds, Nigini Oliveira, Katharina Reinecke, Lionel Robert Jr, Kanya Siangliulue, Volker Wulf, and Chien-Wen Yuan. 2019. Learning from Team and Group Diversity: Nurturing and Benefiting from our Heterogeneity. In *Conference Companion Publication of the 2019 on Computer Supported Cooperative Work and Social Computing*. 498–505.

[42] Robert Munro, Steven Bethard, Victor Kuperman, Vicky Tzuyin Lai, Robin Melnick, Christopher Potts, Tyler Schnoebelen, and Harry Tily. 2010. Crowdsourcing and language studies: the new generation of linguistic data. In *Proceedings of the NAACL HLT 2010 workshop on creating speech and language data with Amazon's Mechanical Turk*. Association for Computational Linguistics, 122–130.

[43] Michael Neff, Yingying Wang, Rob Abbott, and Marilyn Walker. 2010. Evaluating the effect of gesture and language on personality perception in conversational agents. In *International Conference on Intelligent Virtual Agents*. Springer, 222–235.

[44] Douglas L Nelson, Cathy L McEvoy, and Thomas A Schreiber. 2004. The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers* 36, 3 (2004), 402–407.

[45] John E Newhagen, John W Cordes, and Mark R Levy. 1995. Nightly@ nbc. com: Audience scope and the perception of interactivity in viewer mail on the Internet. *Journal of communication* 45, 3 (1995), 164–175.

[46] David Novick and Iván Gris. 2014. Building rapport between human and ECA: A pilot study. In *International Conference on Human-Computer Interaction*. Springer, 472–480.

[47] Jay F Nunamaker, Douglas C Derrick, Aaron C Elkins, Judee K Burgoon, and Mark W Patton. 2011. Embodied conversational agent-based kiosk for automated interviewing. *Journal of Management Information Systems* 28, 1 (2011), 17–48.

[48] Magalie Ochs, Catherine Pelachaud, and Gary Mckeown. 2017. A User Perception–Based Approach to Create Smiling Embodied Conversational Agents. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 7, 1 (2017), 1–33.

[49] Changhoon Oh, Jungwoo Song, Jinhan Choi, Seonghyeon Kim, Sungwoo Lee, and Bongwon Suh. 2018. I lead, you help but only with enough details: Understanding user experience of co-creation with artificial intelligence. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–13.

[50] Stephen Oliver. 2019. Communication and trust: rethinking the way construction industry professionals and software vendors utilise computer communication mediums. *Visualization in Engineering* 7, 1 (2019), 1.

[51] OpenAI, Christopher Berner, Greg Brockman, Brooke Chan, Vicki Cheung, Przemysław Dębiak, Christy Dennison, David Farhi, Quirin Fischer, Shariq Hashme, Chris Hesse, Rafal Józefowicz, Scott Gray, Catherine Olsson, Jakub Pachocki, Michael Petrov, Henrique Pondé de Oliveira Pinto, Jonathan Raiman, Tim Salimans, Jeremy Schlatter, Jonas Schneider, Szymon Sidor, Ilya Sutskever, Jie Tang, Filip Wolski, and Susan Zhang. 2019. Dota 2 with Large Scale Deep Reinforcement Learning. (2019). https://arxiv.org/abs/1912.06680

[52] Ted Pedersen, Siddharth Patwardhan, and Jason Michelizzi. 2004. WordNet:: Similarity: measuring the relatedness of concepts. In *Demonstration papers at HLT-NAACL 2004*. Association for Computational Linguistics, 38–41.

[53] Sheizaf Rafaeli. 1988. Interactivity From New Media to Communication', pp110-34 in R. Hawkins et al.(eds) advancing Communication Science: Merging Mass and Interpersonal Processes Newbury Park. (1988).

[54] Michael Rovatsos, Dagmar Gromann, and Gábor Bella. 2018. The Taboo Challenge Competition. *AI Magazine* 39, 1 (2018), 84–87.

[55] Saqib Saeed, Sardar Zafar Iqbal, Hina Gull, Yasser A Bamarouf, Mohammed A Alqahtani, Madeeha Saqib, and Abdullah M Alghamdi. 2019. Collaboration at Workplace: Technology Design Challenges of Segregated Work Environments. In *2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT)*. IEEE, 1–5.

[56] Ameneh Shamekhi, Q Vera Liao, Dakuo Wang, Rachel KE Bellamy, and Thomas Erickson. 2018. Face Value? Exploring the effects of embodiment for a group facilitation agent. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 391.

[57] Weiyan Shi, Xuewei Wang, Yoo Jung Oh, Jingwen Zhang, Saurav Sahay, and Zhou Yu. 2020. Effects of Persuasive Dialogues: Testing Bot Identities and Inquiry Strategies. *arXiv preprint arXiv:2001.04564* (2020).

[58] David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, and others. 2018. A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science* 362, 6419 (2018), 1140–1144.

[59] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, and others. 2017. Mastering the game of go without human knowledge. *Nature* 550, 7676 (2017), 354–359.

[60] Ji Hee Song and George M Zinkhan. 2008. Determinants of perceived web site interactivity. *Journal of marketing* 72, 2 (2008), 99–113.

[61] Richard S Sutton, David A McAllester, Satinder P Singh, and Yishay Mansour. 2000. Policy gradient methods for reinforcement learning with function approximation. In *Advances in neural information processing systems*. 1057–1063.

[62] Karolis Tijunaitis, Debora Jeske, and Kenneth S Shultz. 2019. Virtuality at work and social media use among dispersed workers: Promoting network ties, shared vision and trust. *Employee Relations: The International Journal* 41, 3 (2019), 358–373.

[63] Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, and others. 2019. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature* 575, 7782 (2019), 350–354.

[64] Oriol Vinyals, Timo Ewalds, Sergey Bartunov, Petko Georgiev, Alexander Sasha Vezhnevets, Michelle Yeo, Alireza Makhzani, Heinrich Küttler, John Agapiou, Julian Schrittwieser, and others. 2017. Starcraft ii: A new challenge for reinforcement learning. *arXiv preprint arXiv:1708.04782* (2017).

[65] Dakuo Wang, Justin D Weisz, Michael Muller, Parikshit Ram, Werner Geyer, Casey Dugan, Yla Tausczik, Horst Samulowitz, and Alexander Gray. 2019. Human-AI Collaboration in Data Science: Exploring Data Scientists' Perceptions of Automated AI. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–24.

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

[67] Ludwig Wittgenstein. 2009. *Philosophical investigations*. John Wiley & Sons.

[68] Yang Xu and Charles Kemp. 2010. Inference and communication in the game of Password. In *Advances in neural information processing systems*. 2514–2522.