Human-AI Collaboration in a Cooperative Game Setting: Measuring Social Perception and Outcomes

ZAHRA ASHKTORAB, IBM Research AI, USA
Q. VERA LIAO, IBM Research AI, USA
CASEY DUGAN, IBM Research AI, USA
JAMES JOHNSON, IBM Research AI, USA
QIAN PAN, IBM Research AI, USA
WEI ZHANG, IBM Research AI, USA
SADHANA KUMARAVEL, IBM Research AI, USA
MURRAY CAMPBELL, IBM Research AI, USA

Human-AI interaction is pervasive across many areas of our day to day lives. In this paper, we investigate human-AI collaboration in the context of a collaborative AI-driven word association game with partially observable information. In our experiments, we test various dimensions of subjective social perceptions (rapport, intelligence, creativity and likeability) of participants towards their partners when participants believe they are playing with an AI or with a human. We also test subjective social perceptions of participants towards their partners when participants are presented with a variety of confidence levels. We ran a large scale study on Mechanical Turk (n=164) of this collaborative game. Our results show that when participants believe their partners were human, they found their partners to be more likeable, intelligent, creative and having more rapport and use more positive words to describe their partner's attributes than when they believed they were interacting with an AI partner. We also found no differences in game outcome including win rate and turns to completion. Drawing on both quantitative and qualitative findings, we discuss AI agent transparency, include design implications for tools incorporating or supporting human-AI collaboration, and lay out directions for future research. Our findings lead to implications for other forms of human-AI interaction and communication.

CCS Concepts: • **Human-centered computing → HCI design and evaluation methods**; **Natural language interfaces**; *Interactive systems and tools*; *Empirical studies in interaction design*; User studies; User interface design.

Additional Key Words and Phrases: Games, AI, Agents, Collaboration, Social Perception

ACM Reference Format:
Zahra Ashktorab, Q. Vera Liao, Casey Dugan, James Johnson, Qian Pan, Wei Zhang, Sadhana Kumaravel, and Murray Campbell. 2020. Human-AI Collaboration in a Cooperative Game Setting: Measuring Social Perception and Outcomes. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW2, Article 96 (October 2020), 20 pages. https://doi.org/10.1145/3415167

Authors' addresses: Zahra Ashktorab, IBM Research AI, Yorktown, NY, USA, zahra.ashktorab1@ibm.com; Q. Vera Liao, IBM Research AI, Yorktown, NY, USA, vera.liao@ibm.com; Casey Dugan, IBM Research AI, Yorktown, 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, Yorktown, NY, USA, zhangwei@us.ibm.com; Sadhana Kumaravel, IBM Research AI, Yorktown, NY, USA, sadhana.kumaravel1@us.ibm.com; Murray Campbell, IBM Research AI, Yorktown, NY, USA, mcam@us.ibm.com.

2573-0142/2020/10-ART96 $15.00
https://doi.org/10.1145/3415167

1 INTRODUCTION

As AI is becoming more pervasive in today's society, it is being used more in collaborative settings, replacing work a human collaborator may have been used for in the past. This includes customer support applications where chatbots are being increasingly employed, for example in empowering low health-literacy patients in hospitals using virtual nurse agents[8], pedagogical systems for tutoring [24], or even creative endeavors such as collaborative writing [16] or art [37]. In computer-mediated communication in which collaboration is taking place between people, HCI researchers have done much work on determining what can lead to many successful collaborations and outcomes, including distance, diversity or inter-generational challenges [25, 39, 49]. Other factors that have been explored are social perceptions between collaborators [46, 54]. As we move into a world of Human-AI collaboration, these perceptions and how they impact the outcomes of the collaboration are not yet well understood. In this work, we study the differences between human-AI and *perceived* human-human social perceptions in a collaborative interaction setting.

One opportunity for studying user perceptions of AI agents in cooperative games are Cooperative Partially Observable games. Cooperative Partially Observable games (CPO) require players to cooperate with one another, especially since there are aspects of the game that are only partially observable [32]. AI research has made strides on competitive games, showing improvements for games like chess, checkers, and poker [9, 13]. More recently however, there has been an interest in the investigation of Cooperative Partially Observable games that are AI-driven. These cooperative games require users to coordinate, cooperate and communicate with an AI agent in ways that their actions are deemed interpretable and interpret the actions of the AI agent. [33]. These cooperative games require consideration of *theory of mind*. In this paper, we introduce a cooperative game that requires the user to interpret an AI agent's clues as there is information withheld from the user.

We studied collaboration between humans and AI in the context of a real-time, interactive collaborative word guessing game. Due to the cooperative nature of the game objective (i.e. guessing a target word based on a partner's clues) and the clear notion of success (i.e. win/loss) in this setting, the game afforded us a unique opportunity to study human-AI collaboration. Further, due to the constrained interactions between partners in the game, we were able to study how social perceptions differ when players believe they are playing with AI versus another human. Prior work investigating social perceptions of AI agents was not in a truly collaborative setting. [8, 26, 50, 56]. Our work is not interested in the intentional anthropomorphizing of a system, but the effect of identity disclosure and the effect of that identity on outcomes and social perceptions of AI or human partner in a collaborative environment.

Our findings show that even while the same AI agent plays across all conditions (i.e. during game play the "human" partners interacted just as the "AI" partners interacted), if people perceive they are playing with a human partner, as opposed to AI, and they believed their partners to be human, they find their partner more intelligent, creative, and likeable. This potential "bias" against an AI partner on dimensions of social perception does not seem to effect collaboration outcomes, however. We also report on findings around transparency (i.e. confidence scores) that can lead to more positive social perceptions of AI agents. We discuss AI agent disclosure and transparency, share design implications from these findings and believe our research leads to new insights and topics for researchers to study how Human-AI collaboration is similar to and different from Human-Human collaboration and even methodology implications for how to run experiments in this space.

In this paper, we investigate the following research questions:

RQ1 How does a participant's belief in whether they are interacting with an AI or a human affect social perceptions of their partner?

RQ2 How does a participant's belief in whether they are interacting with an AI or a human affect game performance/outcome of the collaboration?
RQ3 How does a partner's confidence level affect social perceptions of their partner?
RQ4 How does a partner's confidence level impact game performance/outcome of the collaboration?

2 RELATED WORK

Our inquiry is motivated by the prevalent use of AI systems in human-AI collaborative environments. Previous research has studied both human-human collaboration through computer mediated communication [19, 29, 31] and human-AI interaction [8, 24]. This research builds on prior work by investigating collaboration between humans and AI in real time, measuring collaboration outcomes, and social perceptions of human/AI partners.

2.1 Cooperative Games with Partially Observable Information

Few studies have looked at the social perception of the agent in the context of a cooperative partially observable (CPO) game [61]. Liang et al. investigated implicature communication in Hanabi, a cooperative partially observable game [36] that has been studied in AI literature as CPO [33]. 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. This paper is one of the first to begin exploring interactions between users and agents in cooperative partially observable games. In this study, we explore social perception of an AI agent and outcome of a game. There have been many AI-infused word association games developed, resulting in studies on how users understand, interact and communicate in these games [62]. Games like the 'Taboo' word game [48] "forces agents to speculate about their partner's understanding of the domain, rather than just performing inference on their own knowledge", and similarly for "Hanabi" [4]. 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, 17, 27]. 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 [63].

2.2 Human-AI Collaboration

The term "human-AI collaboration" has emerged in recent work studying user interaction with AI systems [2, 11, 45, 58]. This marks both a shift to a collaborative instead of automated perspective of AI, and the advancement of AI capabilities to be a truly interactive and collaborative partner in some domains. For example, utilizing deep learning technologies that can reach human comparable performance in drawing, Oh et al. explored user experience in drawing pictures collaboratively with AI [45]. They found that while people perceived the AI to be low in predictability, controllability, and comprehensibility, they enjoyed the collaborative experience. Wang et al. surveyed data scientists on their attitudes toward AI systems automating data science projects and found they remain optimistic about a collaborative future with such technologies [58]. A collaborative framework has also been used to study AI-assisted decision systems to understand how to improve human-AI joint decision outcomes [65], users' onboarding needs to facilitate the collaboration [10] and the appropriate level of human and AI contributions in a collaboration [30, 38].

In this paper, we study cooperative games as a new domain of human-AI collaboration. Games are both an important application domain of AI and frequently used as a test bed for state-of-the-art AI algorithms [13, 23, 53, 55], leading to current game-playing AI reaching human-comparable

performance. Meanwhile, the user experience of gaming includes unique aspects to consider for a collaborative partner. One of the aspects we focus on in this paper is the *social perception*, the impression people form about the AI, especially in terms of human-like social characteristics. Unlike purely utility tools, such as decision-support AI, game-playing AI applications often adopt anthropomorphic designs or even disguise as human players [18]. Such designs are not only of hedonic value but could also affect gamer behaviors, such as cooperative attitudes and decisions with AI [28].

Outside collaborative and gaming contexts, research on embodied conversational agents and robots have focused on how the embodiment and anthropomorphic design of agents (not necessarily AI-powered) affect people's social perception. Prior work investigated how people anthropomorphize and attribute social characteristics to robots in general [21]. For example, multiple scales were developed to measure people's self-reported social perception of robots [5, 14] including likability, perceived intelligence, warmth and competence. A bulk of literature studied the effect of embodiment, e.g., adding a talking head or full-fledged virtual body to a system, on people's social perception and task performance for pedagogical tools [8, 34], soliciting questionnaire response [56], autonomous driving [26], group decision support [50] and so on. A meta-analysis of literature on the topic concluded that embodiment enhances social perception but the effect on task performance is small [64]. Research also explored other anthropomorphic design of agents such as personality [35], and using emoji [6].

Our work differs from prior work in that we are not interested in design that intentionally anthropomorphizes computing systems, but the effect disclosure of AI-identity, for an AI partner that has human-comparable performance, has on people's social perception and performance. A few recent works explored the effect of disclosure. 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 [28]. Shi et al. found that compared to a human identity, people are less persuaded by a chatbot even when the same dialogue are used [52]. Interestingly, 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.

Our work is also informed by CSCW and HCI research that explored the factors that can lead to successful collaboration outcomes through computer mediated communication [25, 39, 49], including investigating social perceptions between collaborators [46, 54]. CMC and AI-MC research informs us that people interpret signals they see online to form impressions about other humans - but what are the signals used to draw conclusions in human-AI collaboration and communication? According to the Hyperpersonal Model [57] humans over-interpret cues in computer-mediated communication to form impressions about their partners. Jakesch et al. coin the term *replicant effect*, that only when individuals do not know whether content has been generated by a human or AI, they mistrust the AI less than the humans [29].

Researchers have been interested in impressions of users towards these intelligent systems and even how they form mental models of these systems [22]. In prior work for example, the mental model of an AI agent is described as having three components: **AI Knowledge**: described as the AI's knowledge of various topics, **Local Behavior**: described as an AI's behavior of an individuals output, and **Global Behavior** described as the AI's behavior at large [22]. In the context of our collaborative word guessing game setting, for example, an example of Local behavior is the memory of prior guesses by an individual within one game, whereas an example of global behavior is the memory of guesses and user behavior across multiple games.

2.3 System Confidence and Expectation Setting

Research on AI systems has focused on explainability and trust. Many studies investigating explainability have explored the role of system transparency, particularly user-perception of the system's confidence or certainty in its response or output. Furthermore, many systems can set expectations by showing AI confidence levels to users. Prior research has shown that setting expectations impacts user satisfaction towards technologies [15, 59, 60]. Kocielnik et al. introduce the accuracy indicator, a chart that communicates the accuracy of the AI. They find that including an accuracy indicator lowers user expectations with respect to the system performance. A number of theories regarding expectation setting have been proposed [7, 15]. One theory that has been tested in many HCI studies is the Expectation Confirmation Model (ECM) [7], which suggests that user satisfaction is directly related to user expectations of the system. Specifically, if a user expects more than the system is capable of delivering, the user will ultimately not be satisfied with the system.

3 SYSTEM DESIGN OF WORDGAME: 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 Wordgame. In Wordgame, 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". The player gets 10 attempts to guess before they lose. If the player inputs the correct word, they win. Figure 1 shows a typical round. Wordgame is cooperative, meaning partners work together for the "guesser" to correctly guess the target word. The cooperative nature of this game means that partners are open and honest in achieving a shared goal.

3.1 AI Agent Description

The AI used to play the Wordgame was developed by a team of researchers. The researchers developed two AI agents, one for the giver role (AI has the target word and provides hints to the user) and one for the guesser role (user has the target role and provides hints to the AI). Below, we describe the high-level technical details of the AI agent. The Giver AI generates candidate hints using free association norms, word embeddings, and WordNet [47] (a collection of word level features like antonyms, synonyms, hypernyms etc). Hints are scored 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. Candidate guesses are generated using free association norms, word embeddings, and WordNet, and scores them based on a GBM (supervised machine learning) model trained on free association norms (as hints). On receiving a hint, the guesser 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 giver agent uses a secret word 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 words from Taboo cards as hints and outputs the candidate with highest score as next clue. Upon receiving a new guess, the giver 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. The AI response typically takes 2-3 seconds. Across all conditions we added a delay of 2 more additional seconds. Wordgame is a collaborative game because users have to interpret the clues given by the AI to guess the target word. Building cooperative games with imperfect information

is a "new frontier for AI research" and requires an elevation of reasoning about the beliefs and intentions of other agents, requiring a consideration of *theory of mind* [4].

In this paper, we approach our research questions in the context of the AI agent playing the giver role and the participants in the study playing the guesser role (AI agent gives hints and participants submit guesses for the target word). The two roles (giver and guesser) have slightly different actual and intended behaviors, so we focus on one role to answer our research questions and conduct our experiment.

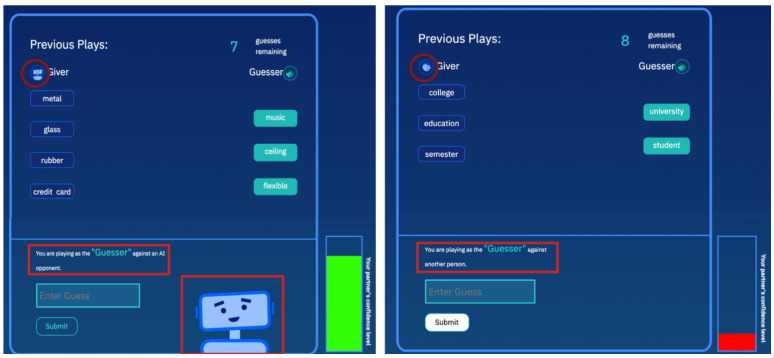


Fig. 1. Game interface detail. On the left, the target word is "plastic" in the consistently high confidence condition. The participant is told they are playing against an AI. On the right, the target word is "school" In this condition, participant is told they are playing against another person in the consistently low confidence condition. Red squares in each interface delineate differences (giver avatar, robot image, and prompt about partner).

4 METHODOLOGY

To better understand how confidence and whether users believed they were playing with an AI agent or a human impacted how they characterized their partner, we ran a large-scale on study on Amazon Mechanical Turk. For this study we limited participants to playing as the guesser and the AI agent to play as the giver.

Each participant played 5 rounds with five different words. For each round, participants were allowed a maximum of 10 guesses. If they did not 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 [22]. 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 were told their partner was: an AI agent, a human, or undisclosed
- Their opponent's confidence: consistently high, consistently low, none displayed, or mixed

Participants were either told they were playing with an AI agent, they were playing with another human, or they might be partnered up with a human or an AI agent. Before being assigned to their opponent, they saw a configuration page that informed them of the nature of their opponent (See Figure 2). Participants were also assigned conditions related to confidence transparency. They were either shown consistently high confidence (75-100) shown in green on every turn, consistently low confidence (0-25) shown in red on every turn, mixed confidence across the turns, or no display of confidence. Participants were told that confidence is a self-report of how confident their gameplay

partners (AI, human, undisclosed) are of each clue. Every time a new clue was given by the agent, a new confidence value within the interval for that condition (consistently high, consistently low, or mixed) was communicated to the user. The colors communicated are colors typically used to express positive settings (high confidence = green), negative settings (low confidence= red) [41]. One reason for using this color scheme was to ensure that participants noticed the variations in the interface so that we could accurately measure the conditions to which they were assigned. The robot avatar appearing during the AI agent condition meant to also more clearly communicate that the user was interacting with a bot and not a human. Two variations of a partner's confidence level are shown in in Figure 1. This was a between-subjects study. Upon registering their Amazon worker ids participants could no longer participate in the HIT. There is some degree of deception in the study and we wanted to prevent the same player participating in multiple conditions as to make the experimental design as reliable as possible.

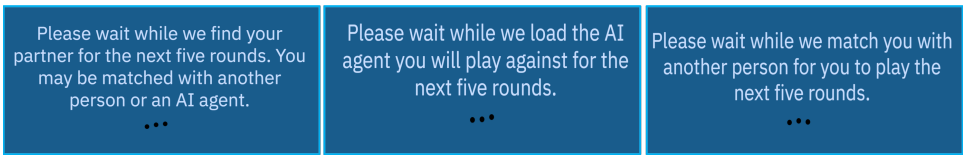


Fig. 2. Configuration page. Participants were told they were either playing with an AI, with another person, or that they may be matched with a human or an AI.

For all players, we used one word list of five words and balanced it for difficulty: "school", "music", "plastic", "ring", "sun". Similar prior work [22] used a similar metric (accessibility index of words, a measure from [42]) to balance for word difficulty. Gero et. al compared user mental models of AI agents in a collaborative word game to win rate. Using two different sets of words balanced for word difficulty, they found similar results. 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).

Participants played a game of 5 rounds and they took a survey (described in the following section). In pilot studies, the average time of completion was 15 minutes. Based on this all participants were paid $3.00 commensurate with federal minimum wage.

4.1 Data Validation

We performed several attentiveness tests to preserve the integrity of the data collected. We excluded those who did not pass the linguistic attentiveness task [40] and workers whose time for finishing the survey portion of the task was less than 15 seconds. We also excluded those whose average ratings for trustworthiness, rapport, creativity, and intelligence fell outside the mean ±2D statistic of participant averages. This left us with 164 subjects.

5 SURVEY INSTRUMENT

We asked participants about demographics, their prior experience with Artificial Intelligence, as well as their prior experience with word games. To ascertain whether participants felt that they were competing against a human or an AI (though in several conditions we told them explicitly whether their partner was an AI or a human), we also collected an AI score as done in [29], in which we asked participants about whether they believed they were interacting with a human or an AI in two questions on a 7-point Likert scale. The AI score is the average of those two questions. We further asked participants to provide open-ended reasons for their AI score. We also asked

Social Perception Index	Cronbach α	Mean	SD
Intelligence	0.79	5.13	1.49
Rapport	0.95	4.65	1.38
Likeability	0.97	4.61	1.35
Creativity	0.95	4.17	1.09

Table 1. Cronbach Alpha, Mean, and SD for each Social Perception Index, N=164

participants to list three attributes they would use to describe their partner and provide reasons for why they selected those attributes.

5.1 Dependent Variables: User Perception of Opponent

To address our research questions, we assessed user perception of rapport, likeability, intelligence and creativity of the opponent. Based on previous work [43], we asked participants to indicate how much 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, creativity), we used a list of semantic differential scales. We adapted scales on these dimensions by [5, 44, 50]. Participants were asked to rate their opponent on pairs of antonyms (i.e. unfriendly/friendly, unpleasant/likeable, ignorant/knowledgeable). All of the perception questions were asked based on a 7 point Likert scale. The averages for perception dimensions were calculated for analysis. Following past studies [5, 44, 50], we combined the items for an Intelligence/Rapport/Likeability and Creativity index by calculating their mean (see Table 1), reliable and consistent with prior work). 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 [5, 44, 50], 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 [43], 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 [5, 44, 50] 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.
- **Creativity** To measure creativity, we used three semantic differential scales in which participants rated their opponent on a 7 point scale as not funny/funny, not creative/creative, and unique/ordinary.

6 RESULTS

When people interact with a human opponent or they interact with an AI opponent, do they have different social perceptions (rapport, likeability, creativity and intelligence) of their opponent? Five regressions were calculated to predict these measures (intelligence, likeability, creativity, rapport, and gameplay) based on the assigned partner (human, AI, undisclosed) and the confidence of the partner (high,low,mixed,none), AI Score, and demographic variables including education, prior exposure to Artificial Intelligence, and prior exposure to word games. Results can be seen in Table 2.

6.1 AI Score: Perceived Partner Type

Often, 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 mistrust the conditions and do not believe them. Additional metrics must be collected to affirm whether participants believed the conditions or to include during analysis to gain a better understanding of all the results. A study investigating how identities and inquiry strategies influence a conversation's effectiveness used deception and found that participants did not believe the conditions of the study. For example, only 34.3% of participants who were told they were interacting with a human believed they were interacting with a human. The authors used a similar approach employed in this paper, by using perceived identity to model their dependent variable [52]. These results further motivated us to include the AI score as a dependent variable in our models, as done in prior work using deception in disclosure of AI identity [52]. In this paper, we investigate whether a participant *believed* whether they were interacting with a human or an AI impacted gameplay and subjective social perceptions. Thus, it is not enough that we simply look at the conditions we assigned to participants (See Figure 2), since not all participants were convinced that they were competing against an AI or an agent. In fact, only 35% of those individuals who we told were interacting with a human in the study believed they were interacting with a human more than an AI (>4.5 on post-survey 1-7 Likert scale). To account for this, we added the AI score, a continuous independent variable, to the model. 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 keep assigned identity in the model because of the possibility that there may be a "suspicion" effect with the same perceived identity in different manipulated conditions, i.e. people are suspicious when we tell them they are interacting with a human, or when we do not disclose with whom they are interacting. We calculated a linear regression for each of the dependent variables and present the results below.

6.2 Demographics and Previous Experience

We asked participants about their demographics (language, education) and previous experience with AI: "What kind of exposure have you had to Artificial Intelligence (AI)" (1= I don't know what machine learning is. , 7= I have implemented a Machine Learning Algorithm) (M=4.7, SD=1.4), as well as prior experience with word games: "How familiar are you with word games like Scrabble, Taboo, crosswords, etc.?" (1= I never play word games. 7 = I play word games everyday) (M=4.7, SD=1.2). We asked these questions on a 7 point Likert scale. We present other demographic information of our participants in Table 3.

6.3 Subjective Social Perception of Opponent

When people are presented with a human partner or an AI partner, do they perceive one as more intelligent, creative, likeable, and have more rapport with one than the other? We conducted a regression to compare the effects of perceived opponent, confidence (low, high, mixed, none), and assigned opponent as part of the condition (human vs. AI. vs. not disclosed) (see Figure 2) and their interaction on perceived intelligence, rapport, likeability and creativity. We treat intelligence, rapport, likeability, creativity and gameplay win rate as dependent variables in our models and describe the results for each below. The details of the each regression are included in Table 2, with details in the subsections below.

6.3.1 Intelligence. A multiple regression was calculated to predict intelligence based on AI score of the partner, the assigned role (human, AI, undisclosed) and the confidence of the partner (high,low,mixed,none), and demographic variables. A significant regression was found F(17,146)=15.22,

Dependent Variable	Significant Effects	β
Intelligence		
	Role of Agent: Undisclosed	-1.98**
	Role of Agent: Human	-1.46***
	AI Score	0.42***
	AI Score x Role of Agent: Human	0.26*
	AI Score x Role of Agent: Undisclosed	0.33*
	Exposure to AI	-0.22***
R^2=0.64***		
Likeability		
	Prior Experience with Word Games	0.11*
	AI Score	0.51***
R^2=0.75***		
Rapport		
	Role of Agent: Human	-1.08 *
	Role of Agent: Undisclosed	-1.51***
	AI Score	0.37***
	AI Score x Role of Agent: Human	0.20*
	AI Score x Role of Agent: Undisclosed	0.31**
R^2=0.58**		
Creativity		
	Exposure to AI	0.10*
	AI Score	0.34***
R^2=0.65***		
Gameplay Win Rate		
R^2=0.15		

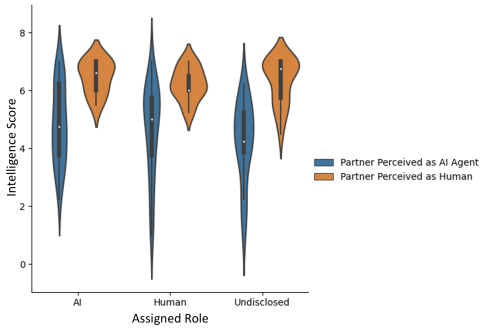
Significance Codes : ***p <0.001, **p <0.01, *p <0.05

Table 2. Regression predicting dependent variables (subjective social perception of opponent) and gameplay win rate based on assigned conditions: assigned role (human, AI, undisclosed), assigned confidence (high, low, mixed, none), AI score, and demographic variables (education, prior experience with word games, age, and prior exposure to artificial intelligence) N=164.

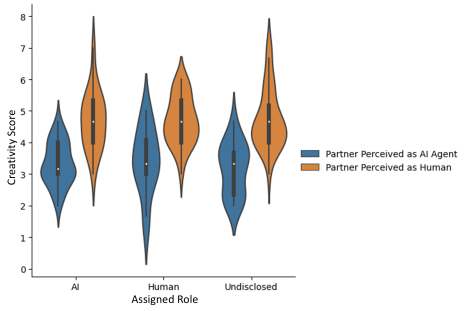
Demographic	
Age	18-25 (12%), 26-35 (42%), 36-45 (24%), 45+ (21%)
Language	English (93%), Tamil (3%), Portuguese (2%), Malayalam (1%), Italian (1%)
Education	High School (33%), Bachelors (58%), Advanced (9%)

Table 3. Participant Demographics, N=164

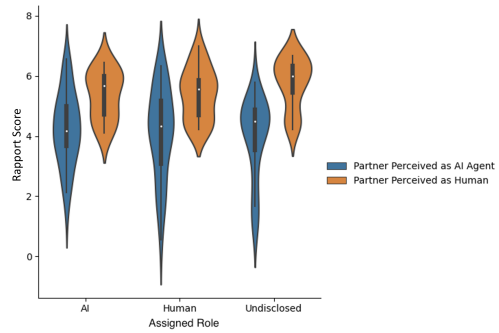
p <0.001, with an R^2 = 0.64. The regression revealed that in the conditions in which participants were told their partners were human, they found their partners to be less intelligent (β= -1.46 ,p <0.001). Similarly, in conditions in which participants were not told definitively whether their partners were an AI agent or a human, participants found their partners to be less intelligent β= -1.98, p <0.01. Conversely, we found a significant main effect for AI score, showing that the more a participant believed that their partner was a human, the more intelligent the partner was perceived (β= 042 ,p <0.001). We also found significant interaction effects between AI score and the Human



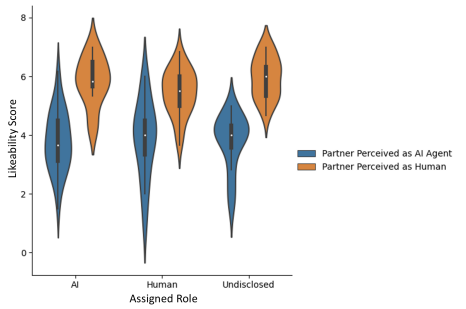
(a) Perceived intelligence of partners measured after cooperative gameplay. Intelligence score is based on the mean of 4 items from the survey (1=Not Intelligent, 7=Intelligent)



(b) Perceived creativity of partners measured after cooperative gameplay. Creativity score is based on the mean of 3 items from the survey (1=Not Creative, 7=Creative)



(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)



(d) 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)

Fig. 3. The social perception measures (intelligence, creativity, rapport, and likeability) plotted against the assigned roles users were given in their conditions. Different hues represent user perception of the agents as AI agent or Human based on the AI score collected at the end of the study.

Agent Role (assigned) condition (β= 0.26 ,p <0.05) and the AI Score and the Undisclosed Agent Role (assigned) condition (β= 0.33 ,p <0.05). These interaction effects show that when individuals believe that they are interacting with a human and we disclosed that they are interacting with a human, or did not disclose at all, they found their partners to be more intelligent.

6.3.2 Likeability. As with intelligence (and all other subjective social perception measures), the regression was calculated based on the AI score of the partner, the assigned role, demographic variables, and the confidence of the partner. A significant regression was found F(17,146) = 22.77, p <0.001 with an R^2 = 0.73. The regression revealed a significant main effect for AI score meaning the more human a partner was perceived, the more likeable it was (β= 0.51, p <0.001). Individuals with more experience with word games also found their partners to be more likeable (β= 0.11, p <0.05). We did not find a significant interaction effect for any of the independent variables.

6.3.3 Creativity. A significant regression was found F(17,146) = 15.75, p <0.001 with an R^2 = 0.65. The regression revealed a significant main effect for AI score, meaning the more human a partner was perceived the more creative it was perceived (β= 0.34, p <0.001). We also found that those with higher exposure to AI found the AI to be more creative (β= 0.10, p <0.01). We also found that there were no significant interaction effects.

6.3.4 Rapport. A multiple regression was calculated to predict rapport based on the AI score of the partner, the assigned role (human, AI, undisclosed) and the confidence of the partner (high,low,mixed,none) and demographic variables. A significant regression was found F(27,146)=11.89, p <0.001, with an R^2 = 0.58. The regression revealed a significant main effect for AI score, showing that the more a participant believed that their partner was a human, the more rapport they reported having with their opponent (β= 0.37, p <0.001). When participants told they were playing with a human agent, they has less rapport with their partners (β=-1.08, p<0.05) . In the undisclosed condition, we found participants also reported having less rapport with their partners (β=-1.51,p <0.001).

6.4 Game Play Results

To investigate **RQ2**, we analyzed the game play results across the conditions with F(17,146) = 1.48, p=0.109 with an R^2= 0.15. We found the model to be insignificant with no main effects or interaction effects found, meaning that AI score, assigned role and confidence did not impact whether participants won or lost games. However, **RQ2** investigates how gameplay was impacted, i.e not only win rate, but also types of words used in response to the opponent. Do participants use different words against an AI versus against a human? Do participants use different words when their opponents' confidence levels are varied?

6.4.1 Unrecognized Words. In Wordgame, unrecognized words or words that are not real words trigger the the "unrecognized word modal" that nudges users to try again. The unrecognized word modal includes the following text: 'To make sure the game runs smoothly, words are checked for spelling before they are sent to you opponent. WORD is not spelled correctly or is not a word. Try again." We recorded the words that triggered the unrecognized word modal in the game, and classify them into two groups: 1) word spelled incorrectly and 2) phrases. Among the phrases used, participants attempted to communicate to their opponent by signaling they did not figure out the target word to win the game or even signaling that they wanted to give up. The unrecognized word modal was triggered 361 times, with the majority (77%) of those words being spelling errors ("dipoloma") or legitimate phrases not recognized by the system ("rice cooker"). 2% of the words were attempts by participants to communicate to their partners including responses like: "i don't know", "hurry up", "i'm not sure", or "i give up".

> *Just tell me what the answer is so we can finish this hit jesus christ.* (Participant #367)

The participant above strongly believed they were interacting with a human as is reflected by their attempt to accelerate completion of the HIT. We compiled the list of users who attempted this sort of communication with their partners. Did they believe they were competing against an AI or a human? Of the 9 participants who employed this sort of tactic to communicate (M=2.8, SD=1.46), the scores for whether they believed they were competing with an AI or human were skewed toward believing that the opponent was a human, which would explain the attempt to communicate with the opponent in this way.

6.4.2 Word Difficulty. We wanted to investigate whether user behavior varied when participants believed they were interacting with an AI or a human. One way to address this question is to take

Perceived Partner	Reason	Example
AI		
	Indirect Clues	*The metal to plastic one was pretty wonky, if it had been credit card first I would think human but I don't think most humans first association to plastic would be metal.*
	Lack of adaptation	*I think a human would have adjusted more to the direction my guesses were going*
	Speed	*Had almost predictable timed responses*
Human		
	High Quality Clues	*I think he is very skilled for giving me good clues*
	Low Quality Clues	*I felt that my opponent was a human. I think an AI Bot would be better at it than a human.*
	Feeling of understanding	*He tried his best to make me understand, his various selection of words been very much helpful to complete the task.*
	Adapting to Guesses	*The opponent seemed to be human, he was able to lead me to correct answers and seemed to adapt his hints to match my guesses somewhat*
	Speed	*The amount of time they took to respond makes me think they were human.*

Table 4. Reasons why participants believed interactions were with an AI partner/Human partner

a closer look at the kinds of words participants used with their partners, including the accessibility index of words, a measure from [42] that reports the frequency that a word is used as opposed to other words. For example, 'cat' will have a higher accessibility index than 'clarinet'. It is related but not identical to frequency of usage. To investigate whether AI score, assigned role and confidence score impacted the kinds of words participants used (i.e. difficult words versus less difficult words) we computed the average accessibility index of the words used during game play and calculated a regression. Our results were not significant and we do not find any differences in the word difficulty when users believed to be competing against an AI or a human or when the partner's confidence score was varied.

6.5 Cues and Signals

We asked participants to explain *why* they believe their partners were AI or human. Two of the coauthors used a grounded theory open-coding approach to extract themes for why users believe they are participating with a human or an AI. Two authors separately extracted codes from the open-ended responses and compared codes, followed by a discussion on codes that were not overlapping. Once all codes were agreed upon, the final list of themes presented in Table 4 was generated, further described there.

6.5.1 AI Opponent. Reasons for why participants believed they were interacting with an AI were because they felt that their opponent was giving **indirect clues**, or clues that were not words that

would be highly associated with the target word. They also felt that the AI **did not adapt** to their guesses and they felt that their opponent's **speed** was too mechanical and quick.

6.5.2 Human Opponent. One theme that emerged from analysis of open responses was around quality of clues. Some respondents justified their answers for believing that their opponent was an human by saying that they felt the clues were **high quality**, while others felt that an AI would give better quality clues than a human.

> *A machine would have given me better answers.* (Participant #360, AI Score = 6)

Participants also noted that they felt that their human partners understood them more and they worked together to solve the puzzle, in that it was a collaborative task. Participants also noted that they felt their opponents adapted to their guesses to steer them in the correct direction, an action that an AI was incapable of.

> *I like how they came up with the words. I kept getting the wrong track (I kept saying soda when they said pop, but they meant pop music, for example).* (Participant #269, AI Score=4.5)

While speed was noted as a reason for believing the opponent was an AI, participants also cited it as a reason for why they believed their opponent was a human.

6.6 Partner Attributes

In addition to the scales of the subjective social perception and game play results, we asked participants to "List three attributes to describe your partner." and "For each attribute listed above, provide a reason." The authors coded the three attributes for sentiment iteratively until all codes were agreed upon. For example, the attribute "unreliable" has negative sentiment whereas "smart" has positive sentiment. 19% of the attributes were mixed/neutral, 22% of the attributes were negative, and 68% of the attributes were positive, resulting in a 9% overlap due to the responses from the participants.

> *I chose thoughtful because he took time to come up with his clue, which means he was thinking. I chose responsive because he tried to play off the words I was guessing to steer me back on track. I chose determined because he didn't give up even though we failed on the first round.* (Participant #257, positive sentiment attributes: "thoughtful, responsive, determined", AI Score=7)

> *There were some answers that did not immediately go with the word so it demonstrated some type of creativity. When my words were wrong they kept working with me and assisting with clues. They tried to think of better words to reach the goal.* (Participant #274, positive sentiment attributes: "creative, persistent, helpful", AI Score =5)

> *Robotic because of the terms it came up with and in what order. Confusing since it started with weird ones such as "metal" when the word was plastic.* (Participant #417 , negative sentiment attributes: "robotic, confusing, limited", AI Score = 2)

6.7 Confidence

Our regression analysis shows that showing a partner confidence or varying it impacted perception of rapport with partner. We found that both in the low confidence condition and no confidence displayed conditions, participants felt more rapport with their partners. The open-ended responses show that participants also considered the confidence meter when judging their partners.

> *The way they changed the kind of clue they gave based on where my interpretation went wrong seemed human. Also, the confidence bar was human-like.* (Participant #530, AI Score=5.5)

Prior work shows that when user expectations are high or do not match the level that the system is capable of delivering, users are less satisfied with their experience [15, 59, 60]. Part of why users had more rapport when the confidence meter was consistently low or when there was no confidence shown to users can be potentially explained by user expectations being lowered by the low confidence meter and then leading to an interaction that yielded a higher rapport score. In fact, one participant described their partner as, "confident, unreliable, clueless", two attributes of which are unquestionably negative, while "confident" when paired with the other characteristics, takes on a negative connotation.

7 DISCUSSION

Our analysis of the AI agent and people's interaction with it during the experimental setup resulted in the following findings. Firstly, people who believe they are interacting with a human - even when the interactions are exactly the same as those interacting with the AI agent, find their partners during game play much more likeable, intelligent, having more rapport, and creative. Further, when describing the attributes of their partners, players used more negative words to describe their partner when they believed they were playing with AI versus more positive words when they believed they were playing with another human. Win rate or the kinds of words participants used with their partners was not impacted when they believed to be interacting with a human or an AI. We also find that users identify specific cues to identify whether they are interacting with an AI or a human that supports prior work of how users form mental models about AI systems. Lastly, we address the limitations of the study and future directions to build on our findings.

7.1 Differences in Human-AI vs Human-Human social perception

Despite our finding that participants found that their *perceived* human partners were much more intelligent, likeable, creative than their AI partners, and even used more negative words to describe their partner when they believed they were paired with AI, we found that this did not impact whether users won/lost game, meaning that having negative perceptions of partners did not lead to a negative outcome of collaboration. So, we can pose the question: Does social perception even matter in human AI-collaboration? In our particular context, we find that outcome is not impacted by negative perceptions of the partner. We know from prior studies of human-human collaborations that social perceptions, especially rapport and trust, yield much better outcomes of collaboration [3, 20]. Future work is needed to compare the similarities and differences between human-human collaboration and human-AI collaboration, especially around further investigating the nuances of different social perception measures on other human-AI collaborative tasks.

7.2 Mistrust in Deception Studies

Consistent with prior work [52], we found that not all of those individuals who were told they were interacting with a human believed they were interacting with human, and not all of the individuals who were told that they were interacting with an AI agent believed they were interacting with an AI agent. Our analyses reveal when we look at the conditions in which we told users that they were interacting with a human, participants actually found their partners to be less intelligent, but if they believed to be interacting with a human in that condition, then they found their partners to be more intelligent. We found a similar phenomenon when measuring rapport. A reason for this could be mistrust of the conditions with which participants were presented with. They simply did not believe the conditions. For this reason, it is important to include a metric that reflects whether participants believed the conditions. In our study, we included the AI score at the end of the study. This question can help researchers understand whether participants believed the conditions and to more accurately interpret their results.

7.3 Signals and Cues: What makes it an AI/Human?

Prior work has discussed various facets of a conceptual model of an AI agent: local behavior, global behavior, and AI Knowledge. Our participant responses (seen in Table 4), show the kinds of signals that participants associated with their AI partners.

7.3.1 When is it an AI?. When describing reasons for why individuals believed they were interacting with an AI agent or a human, people cited the quality of clues. Two clear groups formed in our responses, with some participants saying that they believe a machine would give *better* or higher quality clues than a human while some participants said they believed that a human would give better quality clues than a machine. When we investigate what is meant by "better", we observe that participants who expected an AI to do "better" expected the guesses to be presented in a better order. For example, one participant said that a human should know that the hint "credit card" should precede the hint "metal" for the target word: plastic.

If we consider local behavior as a component of how users form mental models of AI agents, the speed with which the agent interacts with a user is a part of that local behavior. Participants in the study noted speed as both a reason they believed they were interacting with a human or an AI. All partners in the collaborative game were interacting with the same speed. However some felt that because of the speed with which their partner responded they were interacting with a human, while others cited speed for believing they were interacting with an AI. Our results show that yes, people have impressions about the kind of knowledge an AI has and how it behaves, but these impressions range, with some expecting an AI to provide higher quality clues, and some expecting a human to provide higher quality clues, some expecting an AI to be quicker than a human and some expecting a human to be quicker than an AI. Participants also noted "adaptation" as a quality that an AI does not exhibit, but a human does. Those who felt that their partner did not explicitly adapt to their guesses, felt that their partner was an AI, while those who felt that their partner adapted to their guesses felt they were competing against a human. Among all of the cues, adaptation was one that participants associated with a human, having memory of all previous guesses and adapting dynamically based on new guesses provided by their partners.

7.4 Ethical Implications for Human-AI Collaborative Settings

Our study shows that participants believe their opponents to be more likeable, creative, intelligent when they believe they are interacting with a human, even when the *same exact* interaction happens. They even go so far as using negative words when describing the attributes of their partner when they perceive they are playing with an AI versus a human. This shows a clear bias against an AI partner. These results might push toward developers and designers of such systems to not want to disclose to users whether they are interacting with an AI. If the goal is for the user to have a better experience, and users find bots less likeable and intelligent, is it ethically sound to be deceptive? Only one US state requires disclosure to users about the involvement of a bot in commercial use [12]. Building on prior work [51], our results show that disclosing to participants whether the interaction is with an AI or a human (as opposed to not disclosing, as in one of our conditions) leads to participants perceiving their partners as more intelligent.

In a recent study on user perceptions of Airbnb profiles created by AI or humans, Jaskesch et al. coin the term "replicant effect": when respondents rated profiles in which they did not know whether content was generated by an AI or by a human, users rated the profiles they suspected of being written by an AI as less trustworthy [29]. In our analysis on of the effect of assigned role (human, AI, undisclosed), AI score, and confidence(low,high,none,mixed), we find that when we did not disclose to participants whether they were competing against an AI or a human, that they found their partners during game play to be less intelligent. These results expand on this prior

work [29] to show that in a mixed source environment, participants are *suspicious* of being deceived and doubt the intelligence of their partners more so than when they are told they are interacting with an AI or they are told they are interacting with a human. Not only is it an ethically sound decision to avoid ambiguity, but it leads to higher social perceptions in the interaction.

8 LIMITATIONS AND FUTURE WORK

We acknowledge the limitations of this study. First, the collaborative task with which participants are assigned is situated in the context of a game, though some of our conclusions can apply to other forms of human-AI collaboration that are more goal-oriented in the context of non-gaming environments. However, individuals are compensated in this study, so they are extrinsically motivated to complete the HIT (human intelligence task) on Mechanical Turk to be compensated for their time. Thus, they are *collaborating* in a goal-oriented environment with an AI or human (depending on their perception). While another goal-oriented task might have been more closely aligned with kinds of human-AI collaborative tasks that happen in the wild (customer support, pedagogical, etc.), we believe that results of our experiment still hold merit given the goal-oriented nature to ultimately complete and be compensated for collaborating with a partner (human or AI).

9 CONCLUSION

In this work, we investigate user perception towards a partner in an AI-driven game when users are told they are competing against a bot and when users are told they are competing against a human. We found that people find their partners to be more intelligent, more likeable, more creative and overall have more rapport with partners if people believed to be interacting with a human. We also investigated the effect of confidence on social perception and find that users had more rapport with partners who did not show confidence or displayed low confidence. Interactions were constant for all participants despite user perception of whether the partner was an AI or a human, but participants also had more positive sentiment towards their "human" partners. However, unlike in human-human collaboration studies, in our human-AI context, these social perceptions did not impact the collaboration outcomes (i.e. game win/loss). 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, et al. 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] Gloria Barczak, Felicia Lassk, and Jay Mulki. 2010. Antecedents of team creativity: An examination of team emotional intelligence, team trust and collaborative culture. *Creativity and innovation management* 19, 4 (2010), 332–345.

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

[5] 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.

[6] Austin Beattie, Autumn P Edwards, and Chad Edwards. 2020. A Bot and a Smile: Interpersonal Impressions of Chatbots and Humans Using Emoji in Computer-mediated Communication. *Communication Studies* (2020), 1–19.

[7] Anol Bhattacherjee. 2001. Understanding information systems continuance: an expectation-confirmation model. *MIS quarterly* (2001), 351–370.

[8] 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.

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

[10] Carrie J Cai, Jonas Jongejan, and Jess Holbrook. 2019. The effects of example-based explanations in a machine learning interface. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. ACM, 258–262.

[11] 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.

[12] California Governor. 2018. California new Autobot Law. https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201720180SB1001.

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

[14] Colleen M Carpinella, Alisa B Wyman, Michael A Perez, and Steven J Stroessner. 2017. The Robotic Social Attributes Scale (RoSAS) Development and Validation. In *Proceedings of the 2017 ACM/IEEE International Conference on human-robot interaction*. 254–262.

[15] Hao-Fei Cheng, Ruotong Wang, Zheng Zhang, Fiona O'Connell, Terrance Gray, F Maxwell Harper, and Haiyi Zhu. 2019. Explaining Decision-Making Algorithms through UI: Strategies to Help Non-Expert Stakeholders. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 559.

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

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

[18] Jonathan Dodge, Sean Penney, Claudia Hilderbrand, Andrew Anderson, and Margaret Burnett. 2018. How the experts do it: Assessing and explaining agent behaviors in real-time strategy games. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–12.

[19] Nicole B Ellison and Danah M Boyd. 2013. Sociality through social network sites. In *The Oxford handbook of internet studies*.

[20] Stanley E Fawcett, Stephen L Jones, and Amydee M Fawcett. 2012. Supply chain trust: The catalyst for collaborative innovation. *Business Horizons* 55, 2 (2012), 163–178.

[21] Susan R Fussell, Sara Kiesler, Leslie D Setlock, and Victoria Yew. 2008. How people anthropomorphize robots. In *2008 3rd ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 145–152.

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

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

[24] 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.

[25] 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.

[26] Renate Häuslschmid, Max von Buelow, Bastian Pfleging, and Andreas Butz. 2017. Supportingtrust in autonomous driving. In *Proceedings of the 22nd international conference on intelligent user interfaces*. 319–329.

[27] 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.

[28] 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.

[29] 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*. ACM, 239.

[30] Vivian Lai and Chenhao Tan. 2019. On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 29–38.

[31] Cliff AC Lampe, Nicole Ellison, and Charles Steinfield. 2007. A familiar face (book): profile elements as signals in an online social network. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 435–444.

[32] 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.

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

[34] Jamy Li, René Kizilcec, Jeremy Bailenson, and Wendy Ju. 2016. Social robots and virtual agents as lecturers for video instruction. *Computers in Human Behavior* 55 (2016), 1222–1230.

[35] Jingyi Li, Michelle X Zhou, Huahai Yang, and Gloria Mark. 2017. Confiding in and listening to virtual agents: The effect of personality. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces*. 275–286.

[36] 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.

[37] Sissi Liu. 2019. Everybody's Song Making: Do-it-yourself with and against Artificial Intelligence. *Performance Research* 24, 1 (2019), 120–128.

[38] Maximilian Mackeprang, Claudia Müller-Birn, and Maximilian Timo Stauss. 2019. Discovering the Sweet Spot of Human-Computer Configurations: A Case Study in Information Extraction. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–30.

[39] 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.

[40] 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.

[41] KAYA NAz and Helena Epps. 2004. Relationship between color and emotion: A study of college students. *College Student J* 38, 3 (2004), 396.

[42] 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.

[43] 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.

[44] 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.

[45] 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.

[46] 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.

[47] 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.

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

[49] 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.

[50] 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.

[51] 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". In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '20)*. 15.

[52] 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).

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

[54] 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.

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

[56] Janet H Walker, Lee Sproull, and R Subramani. 1994. Using a human face in an interface. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 85–91.

[57] Joseph B Walther. 2007. Selective self-presentation in computer-mediated communication: Hyperpersonal dimensions of technology, language, and cognition. *Computers in Human Behavior* 23, 5 (2007), 2538–2557.

[58] 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.

[59] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y Lim. 2019. Designing Theory-Driven User-Centric Explainable AI. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 601.

[60] Gesa Wiegand, Matthias Schmidmaier, Thomas Weber, Yuanting Liu, and Heinrich Hussmann. 2019. I Drive-You Trust: Explaining Driving Behavior Of Autonomous Cars. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, LBW0163.

[61] Piers R Williams, Diego Perez-Liebana, and Simon M Lucas. 2016. *Cooperative games with partial observability*. Technical Report. Technical Report. IGGI.

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

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

[64] Nick Yee, Jeremy N Bailenson, and Kathryn Rickertsen. 2007. A meta-analysis of the impact of the inclusion and realism of human-like faces on user experiences in interfaces. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 1–10.

[65] Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. 2020. Effect of Confidence and Explanation on Accuracy and Trust Calibration in AI-Assisted Decision Making. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM.

Received January 2020; revised June 2020; accepted July 2020