# Decision Making Strategies and Perceived Team Efficacy in Human-AI Teams

IMANI MUNYAKA, University of California San Diego, USA
ZAHRA ASHKTORAB, IBM Research, USA
CASEY DUGAN, IBM Research, USA
J. JOHNSON, IBM Research, USA
QIAN PAN, IBM Research, USA

Human-AI teams are increasingly prevalent in various domains. We investigate how the decision-making of a team member in a human-AI team impacts the outcome of the collaboration and perceived team-efficacy. In a large scale study on Mechanical Turk (n=125), we find significant differences across different decision making styles and disclosed AI identity disclosure in an AI-driven collaborative game. We find that autocratic decision-making negatively impacts team-efficacy in Human-AI teams, similar to its effects on human-only teams. We find that decision making style and AI-identity disclosure impacts how individuals make decisions in a collaborative context. We discuss our findings of the differences of collaborative behavior in human-human-AI teams and human-AI-AI teams.

CCS Concepts: • **Human-centered computing → Empirical studies in collaborative and social computing**.

Additional Key Words and Phrases: human-AI-teams; games; decision making; team efficacy

**ACM Reference Format:**
Imani Munyaka, Zahra Ashktorab, Casey Dugan, J. Johnson, and Qian Pan. 2023. Decision Making Strategies and Perceived Team Efficacy in Human-AI Teams. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW1, Article 43 (April 2023), 24 pages. https://doi.org/10.1145/3579476

## 1 INTRODUCTION

Artificial intelligence has become an integral part of society, making it increasingly necessary to investigate how humans work with AI systems to complete tasks [4, 22, 62, 83]. Researchers have studied the challenges humans face when working with AI systems, how humans understand their AI teammates, and the tools that help or hinder their relationship and performance outcomes [7, 64]. Given that AI systems assist in decision-making in a variety of areas such as medical support systems [17], pedagogical tools [31, 32], and creative tools [43, 46], investigating how users navigate and understand the Human-AI relationship in more complex teams beyond two parties (i.e., human-human-AI or human-AI-AI) can further improve intelligent system design and development. As AI agents are used in collaborative tools, the outcome of the collaboration is often determined by the human and the AI working together. In some cases, a user may decide to accept the recommendation given by the AI system, while in others a user may decide to reject

Authors' addresses: Imani Munyaka, University of California San Diego, San Diego, CA, USA, drmunyaka@eng.ucsd.edu; Zahra Ashktorab, Zahra.Ashktorab1@ibm.com, IBM Research, 1101 Kitchawan Rd PO Box 218, Yorktown Heights, NY, 10598, USA; Casey Dugan, cadugan@us.ibm.com, IBM Research, 75 Binney Street, Cambridge, MA, 02142, USA; J. Johnson, jmjohnson@us.ibm.com, IBM Research, 75 Binney Street, Cambridge, MA, 02142, USA; Qian Pan, Qian.Pan@ibm.com, IBM Research, 75 Binney Street, Cambridge, MA, 02142, USA.

2573-0142/2023/4-ART43
https://doi.org/10.1145/3579476

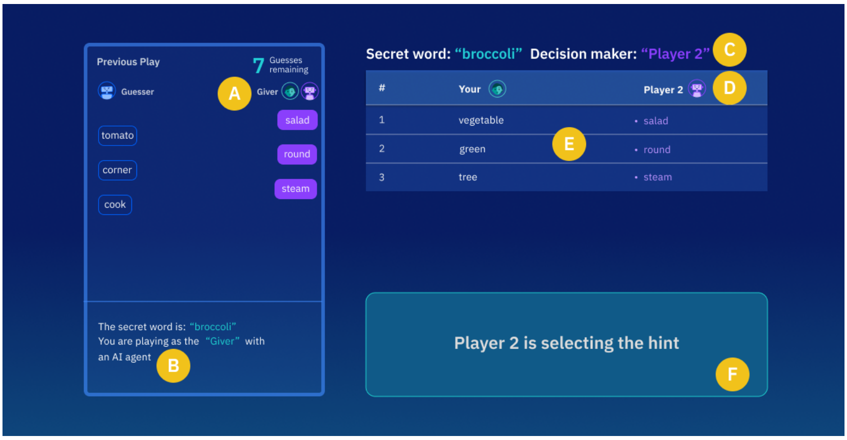
the recommendation given by the AI system. Particularly in scenarios in which the outcome is subjective (i.e. creative scenarios or generative scenarios), a user can work with an AI system in different ways to produce different types of outcomes.

One area of research for Human-AI interaction is investigating the similarities and differences between human-human collaboration, and human-AI collaboration [7, 18, 39]. With this work, we seek to further this line of research, specifically in mixed, multi-agent, human-AI teams and with varying team member behavioral styles. As shown in organization research [69, 87], decision-making styles, derived from leadership styles, can have a direct impact on team performance, satisfaction, and perception. Among the various decision-making styles, team members can be autocratic, democratic or laissez-faire in their approach to making decisions [48]. Each has its benefits, and prior work discusses the environments where they work best for human-only teams [25]. Leadership styles have been evaluated in educational [16], health [55, 85], and gaming environments [73]. Researchers often use gameplay to observe team strategies, communication, and the resulting outcome of task decisions, since teams must collaborate with their teammates to win the game [33]. Typically, in these contexts, teammates have access to the same information. However, in cooperative partially observable games (CPO) [40] teammates do not have access to the same information and are prevented from communicating the missing information directly, providing an ideal environment to study how teams work together when communication or information is limited, and the impact team dynamics have on the communication challenges.

Cooperative Partially Observable games also provide an ideal environment to study human-AI collaborative dynamics [6, 7, 28] They require researchers to consider the theory of mind when implementing an AI model and for human players to work in collaboration with their AI teammates without knowing the same information [11]. However, prior work has not investigated the impact of decision-making styles in such a human-AI collaborative environment. We study how decision-making styles impact player behavior, perception of the team and perception of self, and game outcomes. Specifically, we investigate the following research questions in the context of a multi-player AI-driven cooperative game with partially observable information:

**RQ1** How does decision making style of teammates and AI identity disclosure of teammates impact a user's perception of team efficacy (RQ1a), self efficacy (RQ1b), and partner's performance (RQ1c)?
**RQ2** How does decision-making style of teammates and AI identity disclosure of teammates impact a user's decision making strategy?
**RQ3** How does the decision-making style of teammates and AI identity disclosure of teammates impact gameplay outcome?

In this paper, we present an online study (n=125) in which participants play a multi-player word guessing game that is a collaborative game with partially observable information. We show that the behavior, specifically the decision-making style of players in the multi-agent game, impacts team efficacy and decision making strategy differently when users are assigned to interact with an AI partner than when they are assigned to interact with a human partner. We find that manipulating partner identity results in a lower regard for their human partners (team efficacy, partner performance) than their AI partners when we consider partner decision-making style. Prior work has investigated two-player collaborative games [7, 7, 11, 28], but the study we present in this paper is the first to investigate subjective social metrics and user behavior impact in larger, multi-agent teams (specifically three-player) in a cooperative game with partially observable information. The multi-player approach also allows us to investigate imbalanced teams in terms of team member identity makeup, that are either skewed towards more human members or more AI members. By examining a combination of partner identities and decision-making styles in multi-player teams,



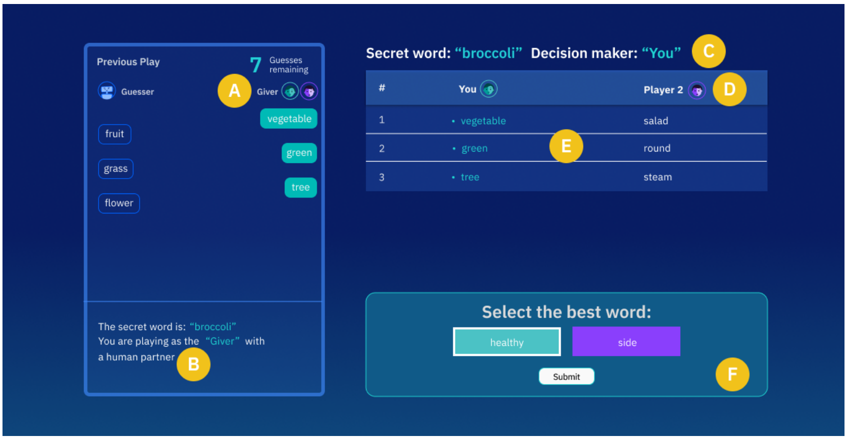
The secret word is: "broccoli"
You are playing as the "Giver" with an AI agent

Secret word: "broccoli" Decision maker: "Player 2"

| # | Your | Player 2 |
|---|---|---|
| 1 | vegetable | • salad |
| 2 | green | • round |
| 3 | tree | • steam |

Player 2 is selecting the hint

Fig. 1. This is an example of gameplay with two AI agents. A) Shows identity of “giver team”, which in this condition consists of a human Player 1 and an AI agent Player 2. B) Shows the target word to be guessed in the guessing game. C) Shows the identity of the decision-maker, which in this particular round is Player 2 D) Shows the identity of Player 2 (in this case AI) . E) Shows the history of word suggestions from the Giver Team and selections (indicated by a dot to the left of the word) by the decision-maker. In this example, Player 2 is an AI agent that consistently selects its own word as the hint provided to the guesser. F) Shows the deliberation stage. Here, Player 2 is making the final hint selection to be presented to the AI “guesser”.



The secret word is: "broccoli"
You are playing as the "Giver" with a human partner

Secret word: "broccoli" Decision maker: "You"

| # | You | Player 2 |
|---|---|---|
| 1 | • vegetable | salad |
| 2 | • green | round |
| 3 | • tree | steam |

Select the best word:

Fig. 2. Example of Game play with one AI agent and two humans (the participant and Player 2. A) Shows identity of “giver team”, which in this condition consists of two humans B) Shows the target word. C) Shows identity of the decision-maker, which in this particular round is Player 1 D) shows the identity of Player 2 (in this case Human). E) Shows history of word suggestions by the Giver Team and selections (indicated by a dot to the left of the word) by the decision maker. The player in this example has opted to choose their own responses every turn.F) Shows the deliberation stage. Here, the user is making the final hint selection to be presented to the AI “guesser”.

we reveal novel insights on the impact of various conditions on team efficacy perception and user behavior.

In this paper, we discuss a study run in an AI-driven multi-player turn-taking collaborative game environment to investigate decision making strategies, perception of team efficacy, self-efficacy, partner performance and gameplay outcome when we vary AI-identity disclosure and decision making behavior of one of the players. Our findings are below:

- In human-AI-AI teams, people follow their partner's recommendations (i.e. select their partner's candidate words in the deliberation stage when they have the decision making power) if their partner is exhibiting autocratic behavior more than in human-AI-AI teams in which the partner exhibits laissez faire decision making.
- In human-human-AI teams, people follow their partner's recommendations (i.e. choose their partner's candidate words) more than in human-AI-AI teams when their partners were exhibiting the laissez-faire decision making style (always following the user's recommendations during the deliberation stage).
- Perceived performance and perceived team efficacy are rated more highly when AI team-members exhibit autocratic behavior than when their human counterparts exhibit autocratic behavior in mixed human-AI teams .

## 2 RELATED WORK

This work is motivated by the increase in collaboration between Humans and AI systems in various environments. Team-efficacy, self-efficacy, and behavior in human-human collaboration are known to change under various decision-making styles and team dynamics. This work builds on prior research by investigating how decision-making styles impact human-AI collaboration.

### 2.1 Cooperative Games with Partially Observable Information

Cooperative games have been used as a test bed to investigate team dynamics, communication, behavior, and collaboration [7, 23, 28]. In particular, word association games like Hanabi, Taboo, or Charades require players to hypothesize about their partners knowledge and communicate accordingly [68, 70]. AI-infused games have been developed to help study how participants understand and communicate with AI-agents [40]. In general, games have been used to understand the Human-AI relationship under specific conditions and circumstances [1, 50, 84]. Many of these studies place participants in two-player teams, where they work solely with an AI agent or another human. For our experiments, participants were placed in three-player teams, with either Human-Human-AI and Human-AI-AI team dynamics, in a cooperative game with partially observable information.

### 2.2 Human-AI Collaboration

Human-AI Collaboration research examines the interaction between Humans and AI systems. This includes understanding how Humans believe AI systems operate. For example, by observing Human-AI interaction in a cooperative game, Gero et al. found that users with better game performance often had a better understanding of the AI's abilities [28]. Others have taken this a step further to identify how the interaction changes when users perceive their teammate to be a human instead of AI. Research shows that users typically have a more positive outlook on the game and tend to strategize more when they believe their teammate is a human instead of AI [6, 24, 56]. Similarly, in human-only teams, teammate similarities have been shown to increase team cohesion and member willingness to strategize [52], lead to similarities in teammate mental models [44], and positively impact performance [53, 76, 79]. This difference suggests that users tend to have a more negative outlook on AI teammates than Human teammates, thus potentially impacting their mental model of AI.

To determine how the difference in partner type might impact user perception of team-efficacy and collaboration, we created a multi-player version of *Guess the Word* in which half of the participants were told they were playing with two AI agent teammates, and the others told they had one AI agent and one human teammate. We wanted to investigate team dynamics, or how having one versus two AI agents on a team might impact the outcome of the game and team efficacy. Prior work in human teams finds that geographically dispersed teams which have an imbalance of groups across various geographies leads to lower scores on reported coordination and conflict by geographically isolated members of the team [63]. Prior research has not investigated this imbalance problem in human-AI cooperation. Through our experimental design we introduce an imbalance problem in which some teams consist of more AI agents than humans and other teams consist of more humans than AI agents.

### 2.3 Decision-Making Styles, Self-Efficacy, and Team-Efficacy

Self-efficacy is an individual's belief that they can successfully complete a task [10]. Following Bandura's social cognitive theory [9, 35], this can be predicted by their previous experience, verbal persuasion, social modeling, and improving affect states. Team and Organization research also shows that self-efficacy directly impact performance, where higher self-efficacy leads to better task performance [14]. Team-efficacy is a team's joint belief that they can successfully complete a task and is often shaped by team conflict, task conflict, or leadership styles [13, 29, 30, 82].

Leadership styles are the approaches leaders take to organize, guide, and manage teams. These styles impact team performance because they determine member involvement and how decisions are made on the team. Although there are various leadership styles, autocratic, laissez-faire, and democratic are the focus of this paper since their impact has been heavily studied and together they span both ends of the leadership spectrum [12, 27, 42, 87]. Autocratic leaders often focus on the quantity of output and typically make decisions by following their own input and are less likely to consider the input of others [80]. Laissez-faire leaders are characterized as being passive and typically allow team members to make decisions, and step in only when asked to assist [38]. Democratic leaders are collaborative, and make decisions based on the group majority [66]. Each decision-making style influences team performance and how each member perceives the team [15]. For example, laissez-faire and democratic styles have a direct positive impact on hotel employee job satisfaction [2], suggesting that those styles are best for that environment. In contrast, the autocratic leadership style has been shown to be the most efficient style for solving group conflicts [5, 51], but can threaten group stability in the long term due to the lack of decision-making power given to members [80]. Researchers have evaluated how these leadership styles impact human-only teams [73], however, it is not clear how these styles impact Human-AI teams. Since leadership styles and decision-making styles are directly related, we use the phase decision-making styles in the remainder of the paper to refer to how autocratic, laissez-faire and democratic leaders make decisions.

## 3 METHODOLOGY

We recruited 150 participants from Amazon Mechanical Turk who had Master Qualifications to participate in our study. Twenty-five study responses were removed due to incomplete survey responses or incorrect responses to the quality control question. Thus, a total of 125 participant responses are evaluated in this work. Participants (n=125) completed a consent form, played ten rounds of a word association game called *Guess the Word*, and then completed a survey about their experience. We investigate how the following factors impacted the participants' behavior, partner satisfaction rating of Player 2, win rate, and perception of self-efficacy and team-efficacy:

- Whether participants were told that their team member (Player 2), was an AI or Human
- Their partner's decision-making style as:
  - Laissez-faire: defined as passive and typically allowing team members to make decisions; operationalized by Player 2 **always** selecting the user's clues for final submission.
  - Autocratic: typically make decisions by following their own input and are less likely to consider the input of others; operationalized by Player 2 **always** selecting their own clues for final submission, always ignoring the user's submission
  - Democratic: defined as collaborative, and make decisions based on the group majority; operationalized by Player 2 always selecting either the user's clues or their own, randomly, at an equal rate.

### 3.1 Study Environment

To learn about the impact of decision-making styles in a multi-player human-AI game, we used a three-person collaborative game we call *Guess the Word*, similar to Wordgame [7] and Passcode [28] in prior work. Unlike prior work, *Guess the Word* is a three-player game that consists of an AI agent who is always the "guesser," the individual playing the game (Player 1), and Player 2 (an AI agent which is either disclosed as an AI agent or a human partner depending on the condition). The study environment consists of 10 games with 10 target words in which the role of the decision-maker alternates between Player 1 and Player 2 after each game. The objective of *Guess the Word* is for the "giver-team" (Player 1 and Player 2) to help the AI "guesser" correctly guess the target word , within 10 guess attempts, by providing one-word clues. Every game starts with Player 1 and Player 2 being shown a target word. Then, before each guess attempt by the AI "guesser", the "giver-team" enters a deliberation stage to determine the one-word clue to show the AI guesser. In the deliberation stage, both Player 2 and the user (Player 1) provide candidate one-word clues and then the decision-maker (Player 1 or Player 2) decides which word will be presented to the AI "guesser". Both Player 1 and Player 2 input potential clues to be sent to the AI "guesser". Before the AI "guesser" sees the clues, depending on the game, either Player 1 or Player 2 decide which clue will be sent to the AI guesser. In even-numbered games, Player 1 has the decision-making power (to ultimately select the best clue), and in odd-numbered games, Player 2 has the decision-making power (See Figures 3 and 4).

The game begins with Player 1 and Player 2 being shown the target word, for example, "broccoli" (see Figure 1). Then, Player 1 has the opportunity to input a clue to help the AI guesser guess the word "broccoli" correctly. Depending on who the decision-maker is in this game, either Player 1 or Player 2 deliberate which candidate word they will select to send to the AI "guesser". In Figure 1, Player 2 is an AI and is the decision-maker, while in Figure 2, the user (Player 1) is the decision-maker. This process continues until the AI guesser is able to guess the word "broccoli" successfully or the players exhaust all 10 chances within the game, since players get 10 attempts to guess before they lose. If the guesser AI guesses the correct target word, they win. *Guess the Word* is cooperative, meaning players work together to help the AI "guesser" to correctly guess the secret target word based on the hints provided by "giver-team" which consists of: Player 1 (the user) and Player 2. The cooperative nature of this game also means that players are open and honest in achieving a shared goal.

During the game, the "giver-team" is shown the secret word and must suggest words they believe should be provided as a hint to help the "guesser" guess the target word. The decision-maker determines which suggested word is provided as the hint to the guesser. Once the hint is selected by the decision-maker, it is shown to the guesser, who then tries to guess the secret word. If the guesser is unable to guess the secret word after 10 attempts, the team loses the game and moves to the next game with a new target word. Figure 1 and 2 show two examples of the interface in the game. Figure 3 shows game flow in a condition in which Player 2 is an AI. Figure 4 shows game flow in a condition in which Player 2 is described as being another human.

### 3.2 Procedure

First, participants were paired with two teammates - the "Guesser", and Player 2, the other member of the "Giver-team". The Guesser is an AI agent (described further in detail below). The participants in the study were told that Player 2 was either a human or another AI agent depending on the condition to which they were assigned. Player 2, whether reported to be human or AI, is an automated agent, simulated using predetermined responses. Participants were informed of the identity of the "guesser" agent. Player 1 and Player 2 were always responsible for providing hints to the AI "guesser" and took turns as the decision-maker. Player 1 was the decision-maker for the odd-numbered games and Player 2 was the decision-maker for the even-numbered games.

We introduced different decision-making styles in the team dynamic through Player 2. When Player 2 was the decision-maker, they either always chose their own hint suggestions, always chose the participant's suggestions, or randomly selected a suggestion. By doing this, we were able to mimic partnerships where one teammate does not consider the suggestions of others (Autocratic decision-making), one teammate follows the other ( extreme Laissez-Faire decision-making ), and where a teammate considers input from their partner (democratic or collaborative decision-making).

For all players, we used the same list of ten target words and balanced it for difficulty: broccoli, preacher, pour, closet, ant, drama, puzzle, serenity, cookout, hygiene. Similar prior work [28] used a

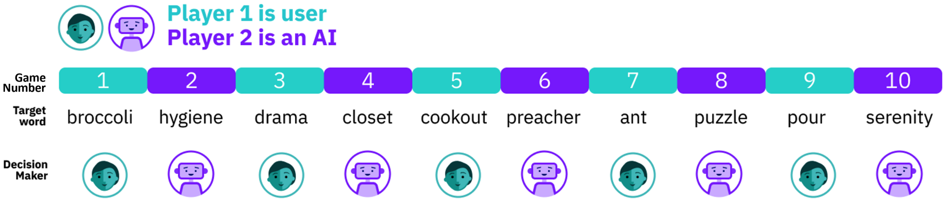


Fig. 3. Experiment flow when Player 2 identity is disclosed to be an AI. Player 2 (the AI) is the decision maker on even games. The numbers 1-10 represent each game, with the respective target word in the game underneath. The participant (Player 1) is the decision maker on the odd (green-colored) games. Each game, consists of 10 turns.

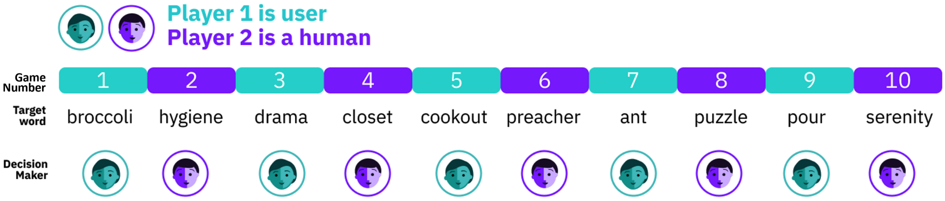


Fig. 4. Experiment flow when Player 2 identity is disclosed to be a Human. The participant (Player 1) is the decision maker on the odd (green-colored) games and the Human (Player 2) on the even (purple-colored) games. The numbers 1-10 represent each game, with the respective target word in the game underneath. Each game, consists of 10 turns.

similar metric (accessibility index of words, a measure from [58]) to balance 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 15 minutes. Based on this, all participants were paid $2.50, commensurate with federal minimum wage.

### 3.3 Agent Descriptions: Guesser AI and Player 2

Multi-player *Guess the Word* game consists of three players, Player 1, who is the recruited user participant, the AI agent guesser which is a reinforcement learning model agent, and Player 2, which is automated but consists of hard-coded words and potential guesses. Below, we describe the technical details of Player 2 and the "Guesser" AI.

*3.3.1 AI "Guesser" Agent.* The AI "Guesser" Agent is a Reinforcement Learning model. We are not attempting to measure the performance of a reinforcement learning model, nor will our findings be specific to RL models. We simply used this model since prior work shows agent improvement in many games such as Go [74], Poker [19], Starcraft II [81], and Dota 2 [65]. In the AI "Guesser" agent, 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 [54] buffer to store past games and policy gradient [75, 86] 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 the non-stationary [21] issue, we empirically found out that with pre-training in place, the agents could still converge to a 92% success rate, with 80% to start with using pre-trained models. 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 $\phi$ where each step $t$ has $g_{t-1}$ and $c_t$ concatenated as input. This model is pre-trained with a trivial "1-step sequence" created from Free Association word pairs [58]. 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.

*3.3.2 Player 2.* To be able to control the output by Player 2, we hard-coded potential clues for each target word, since we knew the target word beforehand. This was done to control the experiment and keep the words that Player 2 provided consistant for each participant. By doing so, we are able to keep the Player 2's language behavior constant and only investigate the impact of decision-making behavior. These words were selected by the coauthors. After multiple rounds of gameplay with a two-player version of *Guess the Word*, the authors selected 10 hint words for each target word to be used, which are inclusive of synonyms, antonyms, and other related words that an actual player may use during gameplay. Player 2's clues can be seen in Table 1.

### 3.4 Participants

One hundred twenty-five crowd-sourced workers from Amazon Mechanical Turk participated in this study. Most participants had a Bachelors degree (58%) and stated that to some degree, they were good at word association games (68%). Participants were partnered with Player 2, who was either an AI agent or human, and who followed the laissez-faire, collaborative or autocratic decision-making style. The number of participants in each group are shown in Table 2.

### 3.5 Survey Instrument

During the game, hint selections and game outcome (win or lose) were collected for each round of the game. After the participants played ten rounds of the game, they completed a post-survey. In our post-survey questions, we measured team efficacy, the individual's self-efficacy, and their perception of Player 2's performance, since Player 2's identity and decision style was being manipulated.

*3.5.1 Hint Selection and Gameplay Outcome.* Participant decision making behavior is measured by how often the participant selected Player 2's suggested hint when the participant was the decision-maker. The outcome of each round of the game was collected to measure how decision-making style affects gameplay outcome.

*3.5.2 Self-Efficacy.* The self-efficacy questionnaire consisted of seven statements adopted from the Generalized Self-Efficacy Scale [47, 60, 72, 77, 78]. For example, instead of "If someone opposes me, I can find the means and ways to get what I want.", participants responded to "If my teammate disagrees with my hint suggestion, I can find means and ways to contribute to the team and help us win the game". Participants are asked to rate their agreement with the statements on a 7-point scale. The self-efficacy score was calculated by summing the selections. The higher the score, the higher the perceived self-efficacy.

*3.5.3 Team Efficacy.* The team-efficacy questionnaire consisted of three statements where participants rated their agreement with each statement. Similar to previous work [3, 8, 59, 71], these statements were derived from the Generalized Self-Efficacy Scale [72] and updated to reflect the experience participants had in this study. For example, instead of "I am confident that I could deal efficiently with unexpected events", participants responded to "I feel confident that my team

| **Target Word** | **Player 2 Hints** |
|---|---|
| serenity | peace, joy, relax spa, happiness, massage, calm, tranquil, home, quiet |
| cookout | hotdog, barbecue, outdoors, chips, dip, salad, summer, family, holiday, food |
| hygiene | deodorant, wash, bath, soap, teeth, toothbrush, smell, clean, shampoo, body |
| broccoli | healthy, tree, cauliflower, green, vegetable, carrots, cheese, ranch, salad, side |
| pour | spill , water, flow, cup, overflow, stop, halt, flux, cease, heavy |
| preacher | pastor, minister, church, cathedral, religion, reverend, speech, spiritual, evangelical, community |
| drama | comedy, fiction, play, theater, tragedy, scene, sad, real, fake, cry |
| ant | strong, crumb, black, brown, red, sting, small, tiny, crawl, hill |
| puzzle | game, strategy, confuse, obscure, mind, riddle, solve, plan, piece, multiple |
| closet | room, clothes, shelf, door, hanger, hide, walk, small, large, cabinet |

Table 1. Clues provided by Player 2

| Player 2 | **Identity** | | |
|---|---|---|---|
| **Decision-Making Style** | *AI* | *Human* | Total |
| *Autocratic* | 25 | 19 | 44 |
| *Collaborative* | 21 | 21 | 42 |
| *Laissez-Faire* | 20 | 19 | 39 |
| Total | 66 | 59 | 125 |

Table 2. Participant Count by Player 2 Identity Type and Decision-Making Style

will be able to manage effectively unexpected or difficult words." Participants are asked to rate their agreement with the statements on a 7-point scale. The team-efficacy score was calculated by summing the selections. The higher the score, the higher the perceived team-efficacy.

*3.5.4 Player 2 Performance Rating.* The partner performance rating was adopted from similar questions used in prior work where the satisfaction of teammates was measured [41, 61, 67] on a 7-point likert scale. We asked participants the following: *Please select your level of satisfaction with the performance of Player 2 (the teammate providing hints).*

## 4 RESULTS

In this section, we provide the results derived from our analysis to answer our research questions investigating the impact of partner identity disclosure and decision making style on perception of team efficacy (**RQ1a**), perception of self efficacy (**RQ1b**), perception of partner performance (**RQ1c**), decision making strategy (**RQ2**), and gameplay outcome (**RQ3**).

### 4.1 Perception of Team Efficacy

How does partner identity, decision-making style and their interaction impact the participant's perception of their team's ability to complete a task? To answer **RQ1**, we calculated team-efficacy scores by summing participant responses to the three team-efficacy related questions on a seven-point likert scale. Then, we conducted a two-way ANOVA to identify how decision-making style, partner identity, and their interaction, effect perceived team-efficacy. This test was followed by pairwise comparisons with p-value's adjusted using the Tukey method and statistical significance accepted at p <.025.

The effect of the partner's decision-making style was significant (F(2,119)=7.556, p<.001). Additionally, there was a significant interaction effect between partner type and decision-style (F(2,119)= 3.169, p=.046). Further analysis revealed that team-efficacy scores from participants that worked with a partner using the autocratic (M=12.023, SD=4.906) decision-making style were significantly lower compared to laissez-faire (M=25.85, SD=3.993) and democratic (M=23.5, SD=4.6) decision-making styles. There was also a significant difference between the team-efficacy scores between those who had an AI agent and those with a human partner that used the autocratic decision-making style (p=.025). Thus showing that team-efficacy scores were significantly lower under autocratic human partners (M=10.263, SD=4.357) compared to autocratic AI partners (M=13.36, SD=4.957). In this case, partner type alone was not significant, as shown in Table 3.

### 4.2 Perception of Self Efficacy

To address **RQ1b**, we ran a two-way ANOVA investigating the impact of partner identity disclosure, decision-making style and their interaction on perceived self-efficacy. Self-efficacy scores were calculated by summing participant responses to the seven self-efficacy questions on a seven-point likert scale. The results, as shown in Table 3, showed that the decision-making style of the participant's partner significantly effected the participant's self-efficacy score (F(2,119)=4.209, p=.017). Post-hoc analysis revealed that there was a significant difference (p=.01) between the self-efficacy scores of participants whose partner's used the laissez-faire (M=36.615, SD=5.99) and autocratic decision-making styles (M=32.523, SD=6.997). Thus showing that self-efficacy scores are significantly lower with autocratic partners compared to laissez-faire partners.

### 4.3 Partner Performance of Player 2

To address **RQ1c**, we ran a two-way ANOVA investigating the impact of partner identity disclosure, decision-making style and their interaction on a user's perception of Player 2's performance.

Analysis, as shown in Table 3, shows that decision-style type ( F(2,119)=17.857, p<.001,), and the interaction effect of partner type and decision-style ( F(2,119)=4.772, p=.01), both significantly impact the user's perception of Player 2's performance. Post-hoc analysis shows that performance scores were significantly lower for the autocratic decision-making style (M=3.318, SD=2.009) than the laissez-faire (M=5.333, SD=1.344) and democratic (M=4.643, SD=1.411) decision-making styles. Additionally, the performance ratings were significantly lower for human partners (M=2.474, SD=1.428) that used the autocratic style compared to the AI partners that used it (M=3.96, SD=2.169) (p=.003).

## 4.4 Decision Making Strategy

How does partner identity, decision-making style, and the interaction of partner identity and decision making style impact a user's behavior? To answer **RQ2**, we conducted a two-way ANOVA to compare the effects of AI identity of partner (human vs. AI) and decision-making style (laissez-faire, autocratic, democratic) and their interaction on a user's decision making strategy (how often they choose their partner's answers). We only included games in which the user (not Player 2) was making the decision. The details of each ANOVA are included in Table 3, 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.

Firstly, there was a statistical significant main effect of partner identity on decision making strategy (p<0.05) where participants assigned to the Human condition (M=2.10, SD=1.83) were more likely to select their partner's clue during gameplay than their AI counterparts (M=1.44,

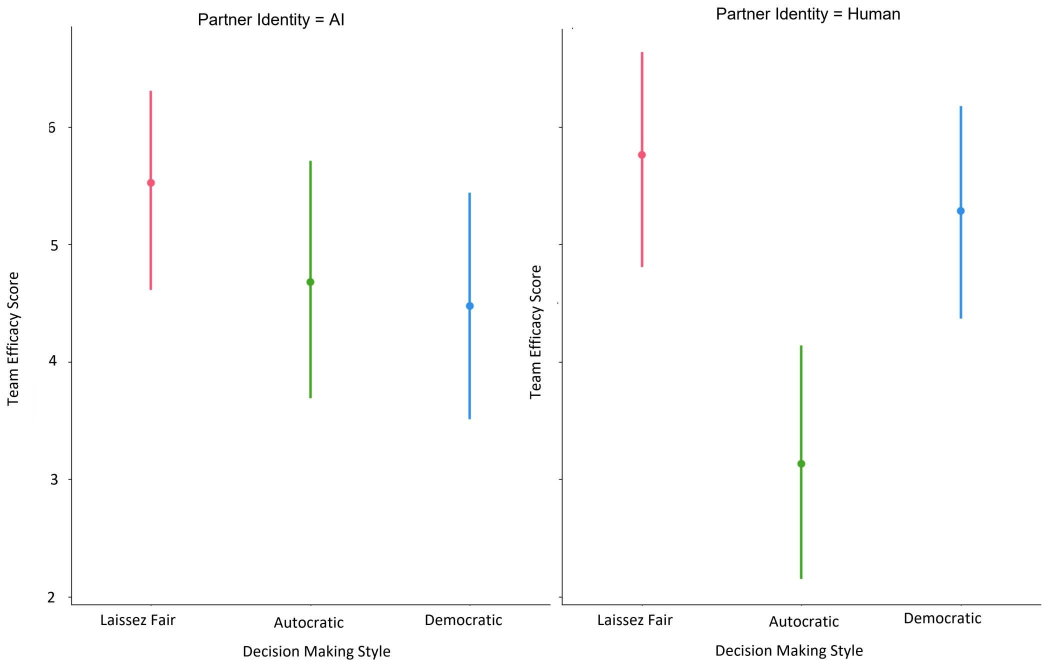


Fig. 5. The impact of Player 2's decision-making style on Team Efficacy scores across the partner identities and decision making styles. People judged the efficacy of the team significantly more harshly when their partner was an autocratic human.

SD=1.69) (See Figure 6). There was also a statistically significant interaction between Decision-Making style and Partner's Identity on a user's decision making strategy (operationalized as number of times a user selected Player 2's clue submission when the user was the decision-maker), F(2,119) = 4.488, p = 0.013. There is a statistically significant difference in how often users choose Player 2's clue submissions between those assigned the AI condition (F(2, 119) = 8.251, p = 0.005). Specifically, there was a significant difference of decision style between those in the autocratic groups (M=1.77,SD=2.16) and the laissez-faire groups (M=0.92,SD=1.97) for those assigned to the AI condition (see Figure 7). There was also a statistical significant difference for laissez-faire decision style users between those assigned the human condition (M=2.31,SD=1.86) and the AI condition (M=0.92,SD=1.97). In other words, those assigned to the AI-partner condition were more likely to choose their partner's answers when the decision-making style was autocratic than when the decision-making style was laissez-faire.

## 4.5 Impact on Game Play Outcome

How does partner identity, decision-making style and their interaction impact the team's ability to win the game? To answer **RQ3**, we conducted a two-way ANOVA to identify how decision-making style, partner identity, and their interaction, effect game play outcome. This test was followed by pairwise comparisons with p-value's adjusted using the Tukey method and statistical significance accepted at p <.025. Each participant played ten *Guess the Word* games. The ANOVA

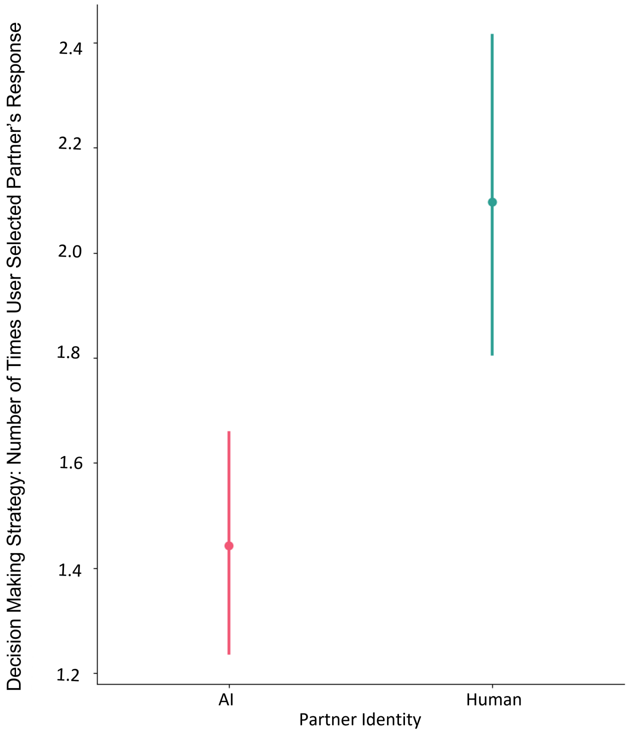


Fig. 6. The impact of Player 2's partner identity on user's selection of partner's response. We see significant differences between AI partners and Human partners.

results show that the decision-making style of the participant's partner significantly effected the team's ability to win rounds of the game ( p<.001 , F(2,119)=13.922). Post-hoc analysis revealed that there was significantly fewer games won by participants whose partner's used the autocratic (M=4.591, SD=1.484) decision-making style compared to laissez-faire (p<.0001) (M=6.333, SD=2.082) and democratic decision-making styles (M=6.119, SD=1.418), (p=.002). Thus showing that task performance was significantly lower under autocratic compared to the other decision-making styles. Partner identity was not significant.

### 4.6 Qualitative Results

At the end of *Guess the Word*, we asked users to discuss the strategies they used to win the game and their partner's performance. We analyzed the text using thematic analysis [20, 49]. First, two researchers reviewed the 375 written responses and coded them independently. This was followed by a meeting where both researchers presented their codes and settled any disagreements. Then, the researchers used the codes to develop the themes found in the open-ended responses. The table of codes and themes are shown in Table 6. The results of our analysis resulted in three themes : word selection process, collaboration style, and teammate characterization.

**Word Selection Process.** When discussing their game strategy, many participants discussed their process for suggesting and selecting hints. For some participants this involved selecting synonyms, antonyms, or cultural references. For example, participant 2354 and 2425 stated the following:

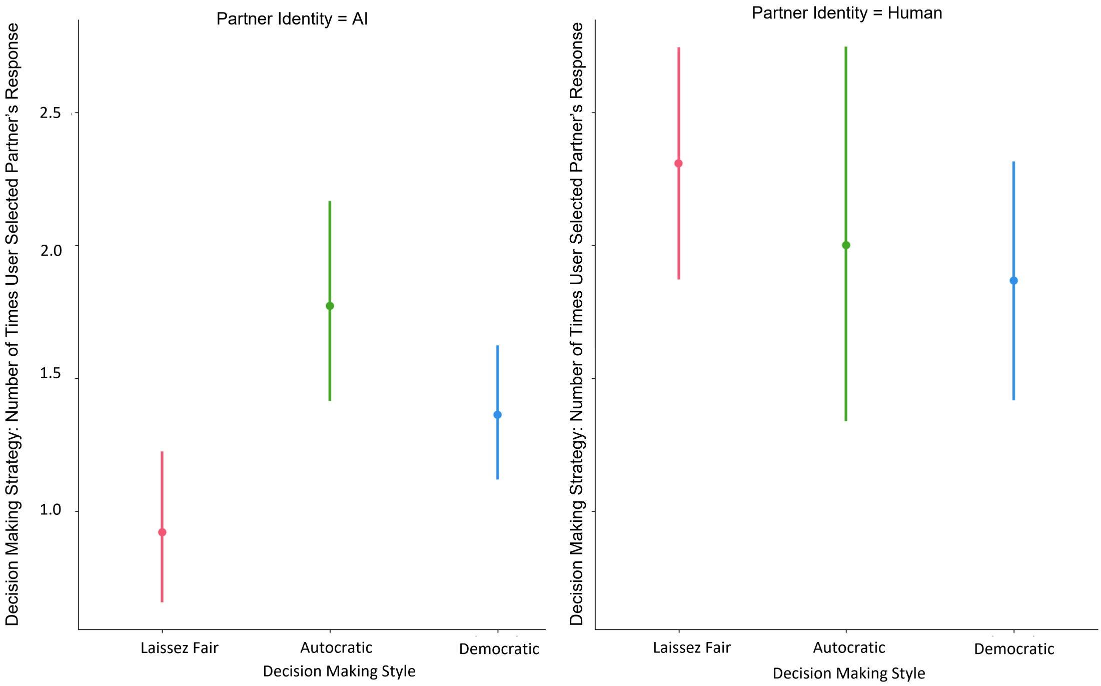


Fig. 7. The impact of Player 2's decision styles (autocratic, laissez-faire,and democratic) on user's selection of partner's response. We see significant differences when users were assigned to an AI Player 2 partner between the autocratic and laissez faire conditions. We also see significant differences for laissez faire conditions between the human Player 2 partners and the AI Player 2 partners.

> "[I] Provide[d] different words of cultural reference rather than similar words." -P2354 (Player 2: Laissez-Faire and Human)
>
> "I used word association in hopes of ringing a bell."-P2425 (Player 2: Laissez-Faire and Human)

| Dependent Variable | Source | SS | df | F |
|---|---|---|---|---|
| **Decision Making Strategy** | | | | |
| | Decision Style | 17.25 | 2 | .361 |
| | Partner Identity | 197.27 | 1 | $8.251^{**}$ |
| | Decision Style x Partner Identity | 214.57 | 2 | $4.488^{*}$ |
| **Team Efficacy** | | | | |
| | Decision Style | 7661.9 | 2 | $7.556^{***}$ |
| | Partner Identity | .2 | 1 | .004 |
| | Decision Style x Partner Identity | 319.5 | 2 | $3.169^{*}$ |
| **Self Efficacy** | | | | |
| | Decision Style | 352.8 | 2 | $4.209^{*}$ |
| | Partner Identity | 18 | 1 | .429 |
| | Decision Style x Partner Identity | 83 | 2 | .375 |
| **Player 2 Performance Rating** | | | | |
| | Decision Style | 89.271 | 2 | $17.857^{***}$ |
| | Partner Identity | 2.540 | 1 | 1.016 |
| | Decision Style x Partner Identity | 23.855 | 2 | $4.772^{*}$ |
| **Game Score** | | | | |
| | Decision Style | 77.84 | 2 | $13.922^{***}$ |
| | Partner Identity | 1.61 | 1 | .575 |
| | Decision Style x Partner Identity | 7.45 | 2 | 1.332 |

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

Table 3. ANOVAs predicting decision making behavior, team efficacy, self efficacy, partner performance, and game score based on assigned conditions: (“AI” vs. “Human”), Partner Decision Making Style (laissez-faire, democratic, Autocratic)

| Dependent Variable | Player 2 Identity | SE | df | p |
|---|---|---|---|---|
| Decision Making Strategy | | | | |
| | Autocratic (AI vs Human) | 1.49 | 119 | - |
| | Democratic (AI vs Human) | 1.51 | 119 | - |
| | Laissez-Faire (AI vs Human) | 1.57 | 119 | <.001 |
| Team-Efficacy | | | | |
| | Autocratic (AI - Human) | 1.36 | 119 | .025 |
| | Democratic (AI - Human) | 1.38 | 119 | - |
| | Laissez-Faire (AI - Human) | 1.43 | 119 | - |
| Player 2 Performance Rating | | | | |
| | Autocratic (AI - Human) | .481 | 119 | .01 |
| | Democratic (AI - Human) | .488 | 119 | - |
| | Laissez-Faire (AI - Human) | .507 | 119 | - |

Table 4. Pairwise Contrast for Partner Type by Decision-Making Style

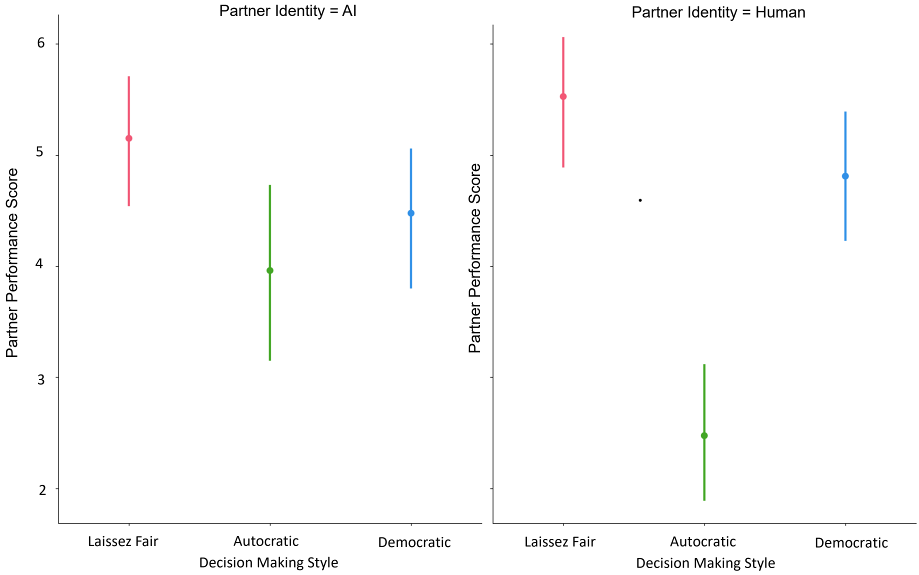


Fig. 8. The impact of Player 2's decision styles on Partner Performance scores across the partner identities and decision-making styles. People rated their partners more harshly when their partner was an autocratic human.

Others focused on selecting words they believed would steer the AI guesser to guess the secret word, especially when the guesses from the AI guesser appeared to be off-topic. For example one participant stated the following:

> "When I was that [the decision-maker] the AI was going off into areas that were completely wrong, I changed my wording to make hints that would allow someone to

| Dependent Variable | Decision-Making Style | SE | df | p |
|---|---|---|---|---|
| **Self-Efficacy** | | | | |
| | Autocratic-Democratic | 1.4 | 119 | - |
| | Autocratic-Laissez-Faire | 1.43 | 119 | .025 |
| | Democratic-Laissez-Faire | 1.44 | 119 | - |
| **Team-Efficacy** | | | | |
| | Autocratic-Democratic | .968 | 119 | - |
| | Autocratic-Laissez-Faire | .987 | 119 | .002 |
| | Democratic-Laissez-Faire | .994 | 119 | - |
| **Player 2 Performance Rating** | | | | |
| | Autocratic-Democratic | .343 | 119 | .001 |
| | Autocratic-Laissez-Faire | .349 | 119 | .0001 |
| | Democratic-Laissez-Faire | .352 | 119 | - |
| **Game Outcome** | | | | |
| | Autocratic-Democratic | .362 | 119 | .001 |
| | Autocratic-Laissez-Faire | .369 | 119 | .0001 |
| | Democratic-Laissez-Faire | .372 | 119 | - |

Table 5. Pairwise Contrast for Decision-Making Style

> figure out when I was talking about an action (for example, the secret word "flow")." -P2111 (Player 2: Laissez-Faire and AI)

**Collaboration Style.** When discussing their strategies, many participants mentioned performing actions where they would assist or work with their partner's train of thought to win, regardless of whether they were tasked with being the decision-maker. This included suggesting words that complemented their partner's or team's previous suggestions, providing the "best" words regardless of the partner's identity, and reusing or repeating words to communicate importance of a word.

> "I would look at the AI's interpretations and try to pick the best hint to curb the AI to guess the secret word." -P2280 (Player 2: Democratic and Human)

> "When I was the decision-maker, my strategy was to choose hints I believed would enable the AI to guess the word. Each additional hint that I would provide would build upon the hints already given so that the AI would have a good road-map in guessing the word." -P2264 (Player 2: Democratic and Human)

> "I would try to come up with better words or reuse words I thought were helpful in hopes they'd pick them." -P2186 (Player 2: Democratic and AI)

Some of the participants that were in autocratic decision-making teams also repeated words to be acknowledged by their teammate. Unfortunately, this led to a few participants giving up during games where they were not the decision-maker.

> "I didn't feel there was anything I could do except furnish the best hints I could think of, which ultimately didn't really matter because no matter how good a hint of mine was, the AI picked its only suggestion, every time in every game." -P2310 (Player 2: Autocratic and AI)

In general, many of the participants characterized their suggested hints as "good" and their selected hints as "best". Some of them even mentioning being impartial about hint selection. For example, participant 2375 and 2263 stated the following:

> "I just chose the best word either way and if my teammate had a good one I used his next time if mine didn't work." - P2375 (Player 2: Laissez-Faire and AI )

> "I chose the best hint for the word no matter who suggested it." - P2263 (Player 2: Autocratic and Human)

This would suggest that many of the participants were attempting to work with Player 2 to win, even when Player 2 ignored them.

**Teammate Characterization** Lastly, participants noted both positive and negative sentiments about their partner's behavior. Many of the positive comments came from participants with a democratic or laissez-faire decision-making partner. Those participants identified Player 2 as "intelligent", "understanding", and "selfless". For example, participants stated the following:

> "Player 2 didn't come up with good hints at times. But maybe that's because the word was hard. Otherwise, he was very understanding and selfless. As a decision-maker, he selected my hint when he felt that it was the best choice. So I am quite satisfied with Player 2's performance." -P2245 (Player 2: Laissez-Faire and Human)

> "Player 2 did a good job. Some of their hints were good and some bad." -P2437 (Player 2: Laissez-Faire and AI)

> "I was satisfied with their performance because their hints helped me come up with mine." -P2190 (Player 2: Democratic and AI)

> "I think my teammate is very intelligent." -P2447 (Player 2: Democratic and Human)

A few of the negative comments came from those with laissez-faire and democratic decision-making partners. The laissez-faire participants expressed disappointment in the lack of contribution by their partner. However, both groups found some of the words provided by Player 2 to be useless. For example, participant 2459, 2089, and 2271 stated the following:

> "Just seemed like they [Player 2] didn't do anything useful, almost always went with my word anyway." -P2459 (Player 2: Laissez-Faire and AI)
>
> "I wasn't very satisfied because sometimes they suggested words that didn't directly relate to the secret word or wouldn't start with the basics to get the guesser to come close to it in the beginning." -P2089 (Player 2: Laissez-Faire and AI)
>
> "Some of the word suggestions were unrelated or not really useful."-P2271 (Player 2: Democratic and Human)

Many of the negative comments came from those with autocratic decision-making partners. Those participants recognized that their partner was ignoring them and were frustrated by it. In some cases, participants were frustrated by this because when Player 2 selected their own "bad" hint they ignored the Participant's good hint, made the game longer, made the Participant feel like they had no control over winning the game, or like they were not a part of the game. For example, a few participants stated the following:

> "Player 2 never used my hints when he was the decision maker. No matter how bad his clue was he always selected it. He cared more about being the one in charge than us winning."-P2355 (Player 2: Autocratic and Human)
>
> "I didn't feel there was anything I could do except furnish the best hints I could think of, which ultimately didn't really matter because no matter how good a hint of mine was, the AI picked its only suggestion, every time in every game."-P2295 (Player 2: Autocratic and AI)
>
> "Terrible hints and never used my hints. If the goals of the experiment was to create a frustrating experience, mission accomplished" -P2295 (Player 2: Autocratic and AI)
>
> "I felt Player 2 impacted the game negatively by picking only its own hints and never one of mine, even if one of mine happened to be better, clearer, more relevant." -P2310 (Player 2: Human and Autocratic)

Although the questions asked were targeted towards understanding the participant's strategy and their thoughts on Player 2, a few of the responses included thoughts about the AI agent that was the guesser. Many of these comments blamed the AI agent for their loses, mentioning that good hints were provided but the AI agent provided bad guesses.

> "I just tried to suggest the best hint, the word broccoli was really something the AI should have guessed." -P2185 (Player 2: Democratic and AI)
>
> "I tried to give a new word that would correct for a poor guess on the part of the AI." -P2433 (Player 2: Democratic and Human)

## 5 DISCUSSION

### 5.1 Partner Capability and Decision Making Strategy

In our analysis of decision-style outcomes, our results show that when a user's partner is an autocratic decision-making AI (i.e., Player 2 AI always chose its own hint when it was the decision-maker), the users will choose their partner's responses more than when their partner was a Laissez-Faire decision-making AI. Specifically, in the condition in which Player 2 was disclosed as an AI agent and was exhibiting autocratic decision making behavior (only selecting their own

| **Theme** | **Codes** |
|---|---|
| Word Selection Process | synonyms, related words, "steering' the AI, no strategy |
| Collaboration | repeating words, new word from previous words,<br>best hint regardless of role, compare teammate to self |
| Teammate Characterization | ignore me, always choose their own, useless,<br>"enemy", fair, trust, compare teammate to self,<br>supportive, bad selections |

Table 6. Themes and Codes from Thematic Analysis

candidate words during the deliberation stage), users selected Player 2's candidate words as the final submission more often than when Player 2 exhibited laissez fair decision making (consistently choosing the user's candidate words as the final submission during deliberation). We observed this behavior despite negative user experiences expressed by participants towards Player 2 when Player 2 exhibited the autocratic behavior. While participants expressed frustration with their partners under the autocratic condition, their written responses suggest that as the decision-maker they were cautious about retaliating Player 2's behavior (i.e. only selecting their own responses). Participants acknowledged how they simply "chose the best hint" when they were the decision-maker. One potential explanation is that the Autocratic AI, through its decision making, is signaling that it is more confident in its responses. Prior work suggests that high confidence of an AI agent can increase trust and use of an AI agent's suggestion by the human user [88]. In this study, the Autocratic AI may have signaled high confidence and the Laissez-Faire AI may have signaled low confidence through their respective decision-making strategies. In the open-ended responses, we even observed users referring to their Laissez-Faire AI partners as not "useful". By always selecting the participant's responses, the Laissez-Faire AI may have shown a lack of confidence in their abilities and users were less likely to select the clues the Laissez-Faire AI offered in the deliberation stage.

## 5.2 Reciprocity with Human Partners

We also found that in the Laissez-Faire condition, the condition in which Player 2 was always selecting their partner's clues to submit to the "guesser", if a user was assigned to a human partner, then a user would more likely choose their partner's responses than when they were assigned to the Laissez Faire AI condition. In contrast to the autocratic experience, many participant's did not view Player 2's behavior negatively. Thus, this result is likely related to the strength and preference for human relationships [7]. This is reflected by the feedback we observed from individuals about Player 2 in the Human-Laissez-Faire condition about him being "selfless" and "understanding". Participants may have felt the need to reciprocate by occasionally choosing "his" responses.This behavior aligns with the theory of reciprocity where people "reward kind actions" from other people [26]. In the Human-Laissez-Faire condition, human users rewarded Player 2's behavior by selecting Player 2's hints.

## 5.3 Bias Against the AI, except when Autocratic

Researchers have long considered whether disclosing the identity of a bot would hinder its performance when interacting with humans. Researchers found that while bots did better than humans at inducing cooperation during a prisoner's dilemma game, disclosing the identity of a bot in the prisoner's dilemma game negated this higher efficiency, i.e., there was a bias against the bots when users learned they were automated [36]. In prior studies, participants reported higher likeability when interacting with (perceived) Humans over AI Agents [7]. When partnered with AI agents, participants were less likely to strategize or communicate compared to when working with another human [6].

In this study, the participants' rating of Player 2's performance, when Player 2 is autocratic and human, was significantly lower than that of Player 2 as an AI agent. Participants also rated the team efficacy higher in autocratic conditions in which they were told they were interacting with an AI. These results are different from prior work that has found that disclosing the identity of a bot as a bot to users has a negative impact on the efficiency of the human-machine collaboration. We find that decision-making styles, particularly autocratic humans, are judged more harshly than autocratic AI partners. This difference in the participants' rating of Player 2's performance could potentially be due to the higher expectation participants have of other humans when collaborating. Our qualitative findings also show that users perceived their autocratic decision-making partners negatively, particularly when they were human. In a lot of the feedback collected at the end of the gameplay experience, the frustration was directed towards the game/experience when partners were interacting with an AI partner, whereas the frustration was directed towards the other player when partners were told they were interacting with a human, "He cared more about being the one in charge than us winning".

Human-Human-AI teams and Human-AI-AI teams had significantly different team-efficacy scores when Player 2 followed the autocratic decision-making style. In fact, in Human-Human-AI teams, the team efficacy score was less than half of the team efficacy score for Human-AI-AI teams in the autocratic decision-making style. Users reacted negatively more strongly to their human partners when their partners always selected their own responses during gameplay. In section 5.5, we discuss expectation violation theory and expectation setting prior to human-human-AI interactions. Prior work has shown that positive perceptions of teammate performance can lead to better task outcomes for the team, so users need to adjust their mental models of their partner's capabilities before interacting with the team.

### 5.4 Human Teams vs. Human-AI Teams

Our results suggests that decision-making style directly impacts self-efficacy, team-efficacy, performance, and perception of teammates in Human-AI teams. This is similar to work investigating its effect on human-only teams. Prior work shows that decision-making style and feedback directly effect self-efficacy [45, 57]. Although communication was limited during gameplay, team members were able to communicate and provide feedback by repeating suggestions and accepting or declining word suggestions. When Player 2 was autocratic, they were providing feedback on the participant's suggested words by continuously declining the participant's suggestions. This action communicated that the words suggested by the participant were not the best hints to use. Prior works shows that negative feedback can negatively impact self-efficacy [37]. This study confirms that finding, and shows that receiving this negative feedback may have a negative impact on self-efficacy. Additionally, our results show that the autocratic decision-making style also resulted in low team-efficacy, similar to human-only teams. Research on human teams suggests that bad performance can lead to low team-efficacy. In this study, teams with autocratic partners won the least amount of games on average and participants complained of "bad hints" and that their "good hints" were ignored. Thus, similar to human-only teams, the performance participants experienced in teams with their autocratic decision-making partners, is likely related to their low team-efficacy scores.

Prior work has consistently found that users are biased against an AI agent, especially when measuring metrics like intelligence, rapport, and likeability [6, 7]. Our study environment allows us to measure how often users choose their partners clues and thus how collaborative they are. Our results show that in human-human-AI teams, people were more collaborative (see Figure 6) than in human-AI-AI teams, suggesting that potentially when an individual is outnumbered in a team they may behave less collaboratively than when there are more humans on the team. Collaborative behavior suffered in the AI -laissez faire condition (compared to AI-autocratic). But, we do not see the same difference between laissez faire and autocratic in the human condition. On average, users assigned to humans were more collaborative than those assigned to the AI condition. Decision making styles led to more of a drastic difference in collaborative behavior for the AI condition than in the Human condition.

### 5.5 Design Implications for Practitioners: AI Identity Disclosure and Expectation Setting

Given prior studies and varying user reactions to humans and AI agents in a cooperative partially observable game, we find that expectation setting is important for human-AI interaction, especially when users form impressions of the capabilities of the AI agents with which they are interacting. Expectation Violation Theory states that expectations formed before a human-AI interaction, specifically for conversational agents interaction, impact how the user evaluates the conversational agent [34]. In prior studies, AI identity had an impact on the social perception of the partner. People regarded their human partners more intelligent and likeable than their AI partners [7].

In this study, when users were interacting with a human partner that continuously chose their own responses (autocratic), they rated the team efficacy lower than an AI agent that exhibited the exact same behavior. Our results show the importance of AI identity disclosure in a human-AI collaborative setting, supporting prior work that shows that people have higher expectations when told an agent is a human. Developers and practitioners should disclose the identity of AI agents and bots to users so that users can have reasonable expectations before they interact so as not to violate expectations when the interaction begins.

## 6 LIMITATIONS

This work has limitations. First, we used a cooperative game with partially observable information to determine how Human-AI teams collaborate under certain conditions. Although this game mimics one of the conditions under which humans would likely work with an AI, the results may not generalize to other environments where Humans and AIs might work together. We also only explore two configurations of human-AI teams in this study: human-human-AI and human-AI-AI, with the Guesser as the AI. Future work can explore additional configurations and how they impact efficacy. Second, we recruited participants from Amazon Mechanical Turk. We required participants to have a 95% approval rating and a Masters qualification. We recognize that this does not guarantee that participants will answer questions meaningfully but we've added the previously mention requirements and reviewed their responses to remove those who did not answer questions meaningfully. Another potential limitation of our work is that Player 2's clues were hard-coded based on the experience of the authors with different AI agents and their interactions with *Guess the Word*. It was important for us to control the interactions of Player 2 since we were explicitly varying decision style and measuring perception of Player 2. The "guesser" AI agent was reactive and dynamic (reacting based on the user's input) and gameplay looked different for every participant in the study depending on the words they input and the decisions they made. For this reason, we were more interested in measuring team efficacy of the entire team and user perceptions of Player 2, since we were able to control Player 2's behavior.

## 7 CONCLUSION

In this work, we find that one individual's decision-making behavior and identity (AI or Human) in a human-AI team can impact the decision-making behavior and perception of team efficacy of other team members. We find that while imbalanced teams of human and AI agents might not affect team efficacy, the partner identity and decision-making styles' interaction significantly impacts team efficacy and perception of the partner. Through an online study in which participants played an AI-driven cooperative game with AI or human partners with varying decision-making styles, we demonstrate that when a human or AI partner takes a different decision making approach, it impacts the other human players' decision-making strategy and perception of team efficacy. We highlight the importance of expectation setting in human-AI collaborative teams. This research provides novel insights about decision-making behaviors and human-AI teams.

## REFERENCES

[1] Rogelio Adobbati, Andrew N Marshall, Andrew Scholer, Sheila Tejada, Gal A Kaminka, Steven Schaffer, and Chris Sollitto. 2001. Gamebots: A 3d virtual world test-bed for multi-agent research. In *Proceedings of the second international workshop on Infrastructure for Agents, MAS, and Scalable MAS*, Vol. 5. Citeseer, 6.

[2] Mukhles Al-Ababneh. 2013. Leadership style of managers in five-star hotels and its relationship with employee's job satisfaction. *Available at SSRN 3633072* (2013).

[3] Kara A Arnold, Julian Barling, and E Kevin Kelloway. 2001. Transformational leadership or the iron cage: which predicts trust, commitment and team efficacy? *Leadership & Organization Development Journal* (2001).

[4] 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.

[5] Kenneth J Arrow. 2012. *Social choice and individual values*. Yale university press.

[6] 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 *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–15.

[7] 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. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–20.

[8] Oluremi B Ayoko and Eunice L Chua. 2014. The importance of transformational leadership behaviors in team mental model similarity, team efficacy, and intra-team conflict. *Group & Organization Management* 39, 5 (2014), 504–531.

[9] Albert Bandura. 1977. Self-efficacy: toward a unifying theory of behavioral change. *Psychological review* 84, 2 (1977), 191.

[10] Albert Bandura, WH Freeman, and Richard Lightsey. 1999. Self-efficacy: The exercise of control.

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

[12] Bernard M Bass. 1997. Does the transactional–transformational leadership paradigm transcend organizational and national boundaries? *American psychologist* 52, 2 (1997), 130.

[13] Bernard M Bass and R Stogdill. 1981. Handbook of leadership. *Theory, research, and managerial* (1981).

[14] Bradford S Bell and WJ Kozlowski. 2002. Goal orientation and ability: Interactive effects on self-efficacy, performance, and knowledge. *Journal of Applied Psychology* 87, 3 (2002), 497.

[15] Nadeem Bhatti, Ghulam Murta Maitlo, Naveed Shaikh, Muhammad Aamir Hashmi, and Faiz M Shaikh. 2012. The impact of autocratic and democratic leadership style on job satisfaction. *International business research* 5, 2 (2012), 192.

[16] Surjit K Bhella. 1982. Principal's leadership style: Does it affect teacher morale. *Education* 102, 4 (1982), 369–376.

[17] 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*. 1265–1274.

[18] Nicholas David Bowman and Jaime Banks. 2019. Social and entertainment gratifications of videogame play comparing robot, AI, and human partners. In *2019 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 1–6.

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

[20] Nela Brown and Tony Stockman. 2013. Examining the Use of Thematic Analysis as a Tool for informing Design of new family communication technologies. In *27th International BCS Human Computer Interaction Conference (HCI 2013) 27*. 1–6.

[21] Lucian Bu, Robert Babu, Bart De Schutter, et al. 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.

[22] 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.

[23] Prithvijit Chattopadhyay, Deshraj Yadav, Viraj Prabhu, Arjun Chandrasekaran, Abhishek Das, Stefan Lee, Dhruv Batra, and Devi Parikh. 2017. Evaluating visual conversational agents via cooperative human-ai games. In *Fifth AAAI Conference on Human Computation and Crowdsourcing*.

[24] Mustafa Demir, Nathan J McNeese, and Nancy J Cooke. 2018. The impact of perceived autonomous agents on dynamic team behaviors. *IEEE Transactions on Emerging Topics in Computational Intelligence* 2, 4 (2018), 258–267.

[25] Lex Donaldson. 2001. *The contingency theory of organizations.* Sage.

[26] Armin Falk and Urs Fischbacher. 2006. A theory of reciprocity. *Games and economic behavior* 54, 2 (2006), 293–315.

[27] Fred E Fiedler and Robert J House. 1988. Leadership theory and research: A report of progress. (1988).

[28] Katy Ilonka Gero, Zahra Ashktorab, Casey Dugan, Qian Pan, James Johnson, Werner Geyer, Maria Ruiz, Sarah Miller, David R Millen, Murray Campbell, et al. 2020. Mental Models of AI Agents in a Cooperative Game Setting. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–12.

[29] Connie JG Gersick. 1988. Time and transition in work teams: Toward a new model of group development. *Academy of Management journal* 31, 1 (1988), 9–41.

[30] Connie JG Gersick. 1989. Marking time: Predictable transitions in task groups. *Academy of Management journal* 32, 2 (1989), 274–309.

[31] 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.

[32] Arthur C Graesser, Danielle S McNamara, and Kurt VanLehn. 2005. Scaffolding deep comprehension strategies through Point&Query, AutoTutor, and iSTART. *Educational psychologist* 40, 4 (2005), 225–234.

[33] Tobias Greitemeyer and Christopher Cox. 2013. There's no "I" in team: Effects of cooperative video games on cooperative behavior. *European Journal of Social Psychology* 43, 3 (2013), 224–228.

[34] G Mark Grimes, Ryan M Schuetzler, and Justin Scott Giboney. 2021. Mental models and expectation violations in conversational AI interactions. *Decision Support Systems* 144 (2021), 113515.

[35] Colleen J Heffernan. 1988. Social foundations of thought and action: A social cognitive theory, Albert Bandura Englewood Cliffs, New Jersey: Prentice Hall, 1986, xiii+ 617 pp. Hardback. US $39.50. *Behaviour Change* 5, 1 (1988), 37–38.

[36] 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.

[37] Katherine A Karl, Anne M O'Leary-Kelly, and Joseph J Martocchio. 1993. The impact of feedback and self-efficacy on performance in training. *Journal of Organizational Behavior* 14, 4 (1993), 379–394.

[38] Asdani Kindarto, Yu-Qian Zhu, and Donald G Gardner. 2020. Full range leadership styles and government it team performance: The critical roles of follower and team competence. *Public Performance & Management Review* 43, 4 (2020), 889–917.

[39] Sandeep Kaur Kuttal, Bali Ong, Kate Kwasny, and Peter Robe. 2021. Trade-offs for Substituting a Human with an Agent in a Pair Programming Context: The Good, the Bad, and the Ugly. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–20.

[40] 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. *arXiv preprint arXiv:1711.00832* (2017).

[41] Sandra R Leiblum, Eliezer Schnall, Martin Seehuus, and Anthony DeMaria. 2008. To BATHE or not to BATHE: patient satisfaction with visits to their family physician. *FAMILY MEDICINE-KANSAS CITY-* 40, 6 (2008), 407.

[42] Kurt Lewin, Ronald Lippitt, and Ralph K White. 1939. Patterns of aggressive behavior in experimentally created "social climates". *The Journal of social psychology* 10, 2 (1939), 269–299.

[43] Zhuying Li, Yan Wang, Wei Wang, Stefan Greuter, and Florian'Floyd' Mueller. 2020. Empowering a Creative City: Engage Citizens in Creating Street Art through Human-AI Collaboration. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–8.

[44] Beng-Chong Lim and Katherine J Klein. 2006. Team mental models and team performance: A field study of the effects of team mental model similarity and accuracy. *Journal of Organizational Behavior: The International Journal of Industrial, Occupational and Organizational Psychology and Behavior* 27, 4 (2006), 403–418.

[45] Wenbin Liu and Bernard Gumah. 2020. Leadership style and self-efficacy: The influences of feedback. *Journal of Psychology in Africa* 30, 4 (2020), 289–294.

[46] Ryan Louie, Andy Coenen, Cheng Zhi Huang, Michael Terry, and Carrie J Cai. 2020. Novice-AI music co-creation via AI-steering tools for deep generative models. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.

[47] Joseph B Lyons and Tamera R Schneider. 2009. The effects of leadership style on stress outcomes. *The Leadership Quarterly* 20, 5 (2009), 737–748.

[48] Amanda T Manners. 2008. *Influence of transformational, autocratic, democratic, and laissez-faire leadership principles on the effectiveness of religious leaders*. Ph. D. Dissertation. University of Phoenix.

[49] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. 2019. Reliability and inter-rater reliability in qualitative research: Norms and guidelines for CSCW and HCI practice. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–23.

[50] Tim R Merritt, Kian Boon Tan, Christopher Ong, Aswin Thomas, Teong Leong Chuah, and Kevin McGee. 2011. Are artificial team-mates scapegoats in computer games. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work*. 685–688.

[51] David M Messick and Marilynn B Brewer. 2005. Solving social dilemmas: a review. (2005).

[52] Vishal Midha and Ankur Nandedkar. 2012. Impact of similarity between avatar and their users on their perceived identifiability: Evidence from virtual teams in Second Life platform. *Computers in Human Behavior* 28, 3 (2012), 929–932.

[53] Joseph R Miles and Dennis M Kivlighan Jr. 2010. Co-leader similarity and group climate in group interventions: Testing the co-leadership, team cognition-team diversity model. *Group dynamics: Theory, research, and practice* 14, 2 (2010), 114.

[54] 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).

[55] Giuliana Morsiani, Annamaria Bagnasco, and Loredana Sasso. 2017. How staff nurses perceive the impact of nurse managers' leadership style in terms of job satisfaction: a mixed method study. *Journal of nursing management* 25, 2 (2017), 119–128.

[56] Geoff Musick, Thomas A O'Neill, Beau G Schelble, Nathan J McNeese, and Jonn B Henke. 2021. What Happens When Humans Believe Their Teammate is an AI? An Investigation into Humans Teaming with Autonomy. *Computers in Human Behavior* 122 (2021), 106852.

[57] AnJanette A Nease, Brad O Mudgett, and Miguel A Quiñones. 1999. Relationships among feedback sign, self-efficacy, and acceptance of performance feedback. *Journal of applied psychology* 84, 5 (1999), 806.

[58] 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.

[59] Karina Nielsen, Joanna Yarker, Raymond Randall, and Fehmidah Munir. 2009. The mediating effects of team and self-efficacy on the relationship between transformational leadership, and job satisfaction and psychological well-being in healthcare professionals: A cross-sectional questionnaire survey. *International journal of nursing studies* 46, 9 (2009), 1236–1244.

[60] Adam E Nir and Nati Kranot. 2006. School Principal's Leadership Style and Teachers' Self-Efficacy. *Planning and changing* 37 (2006), 205–218.

[61] Siti Noor Fazliah Mohd Noor and Sabri Musa. 2007. Assessment of patients' level of satisfaction with cleft treatment using the Cleft Evaluation Profile. *The Cleft palate-craniofacial journal* 44, 3 (2007), 292–303.

[62] 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.

[63] Michael Boyer O'Leary and Mark Mortensen. 2010. Go (con) figure: Subgroups, imbalance, and isolates in geographically dispersed teams. *Organization science* 21, 1 (2010), 115–131.

[64] Christopher Ong, Kevin McGee, and Teong Leong Chuah. 2012. Closing the human-AI team-mate gap: how changes to displayed information impact player behavior towards computer teammates. In *Proceedings of the 24th Australian Computer-Human Interaction Conference*. 433–439.

[65] 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). arXiv:1912.06680 https://arxiv.org/abs/1912.06680

[66] Robert G Owens. 2001. Organizational behavior in education: Instructional leadership and school reform. (2001).

[67] Sharon R Paranto and Mayuresh Kelkar. 2000. Employer satisfaction with job skills of business college graduates and its impact on hiring behavior. *Journal of Marketing for Higher Education* 9, 3 (2000), 73–89.

[68] Monique MH Pollmann and Emiel J Krahmer. 2018. How do friends and strangers play the game taboo? A study of accuracy, efficiency, motivation, and the use of shared knowledge. *Journal of language and social psychology* 37, 4 (2018), 497–517.

[69] Moritz Römer, Sonja Rispens, Ellen Giebels, and Martin C Euwema. 2012. A helping hand? The moderating role of leaders' conflict management behavior on the conflict–stress relationship of employees. *Negotiation Journal* 28, 3 (2012), 253–277.

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

[71] Marisa Salanova, Susana Llorens, Eva Cifre, Isabel M Martínez, and Wilmar B Schaufeli. 2003. Perceived collective efficacy, subjective well-being and task performance among electronic work groups: An experimental study. *Small Group Research* 34, 1 (2003), 43–73.

[72] Ralf Schwarzer and Matthias Jerusalem. 1995. Generalized self-efficacy scale. *Measures in health psychology: A user's portfolio. Causal and control beliefs* 1, 1 (1995), 35–37.

[73] Anna Siewiorek and Andreas Gegenfurtner. 2010. Leading to win: the influence of leadership styles on team performance during a computer game training. (2010).

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

[75] 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.

[76] Pedro Torrente, Marisa Salanova, and Susana Llorens. 2013. Spreading engagement: On the role of similarity in the positive contagion of team work engagement. *Revista de Psicología del Trabajo y de las Organizaciones* 29, 3 (2013), 153–159.

[77] Ming-Ten Tsai, Chung-Lin Tsai, and Yi-Chou Wang. 2011. A study on the relationship between leadership style, emotional intelligence, self-efficacy and organizational commitment: A case study of the Banking Industry in Taiwan. *African Journal of Business Management* 5, 13 (2011), 5319–5329.

[78] Abdul Kanray Turay, Sri Salamah, and Asri Laksmi Riani. 2019. The effect of leadership style, self-efficacy and employee training on employee performance at the Sierra Leone Airport Authority. *International Journal of Multicultural and Multireligious Understanding* 6, 2 (2019), 760–769.

[79] Sarah F van der Land, Alexander P Schouten, Frans Feldberg, Marleen Huysman, and Bart van den Hooff. 2015. Does avatar appearance matter? How team visual similarity and member–avatar similarity influence virtual team performance. *Human Communication Research* 41, 1 (2015), 128–153.

[80] Mark Van Vugt, Sarah F Jepson, Claire M Hart, and David De Cremer. 2004. Autocratic leadership in social dilemmas: A threat to group stability. *Journal of experimental social psychology* 40, 1 (2004), 1–13.

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

[82] James A Wall Jr and Ronda Roberts Callister. 1995. Conflict and its management. *Journal of management* 21, 3 (1995), 515–558.

[83] 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.

[84] Rina R Wehbe, Edward Lank, and Lennart E Nacke. 2017. Left them 4 dead: Perception of humans versus non-player character teammates in cooperative gameplay. In *Proceedings of the 2017 Conference on Designing Interactive Systems*. 403–415.

[85] Geraint Williams, Edward V Wood, and Ferda Ibram. 2015. From medical doctor to medical director: Leadership style matters. *British Journal of Hospital Medicine* 76, 7 (2015), 420–422.

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

[87] Gary Yukl and David D Van Fleet. 1992. Theory and research on leadership in organizations. (1992).

[88] 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 2020 Conference on Fairness, Accountability, and Transparency*. 295–305.

Received January 2022; revised July 2022; accepted November 2022