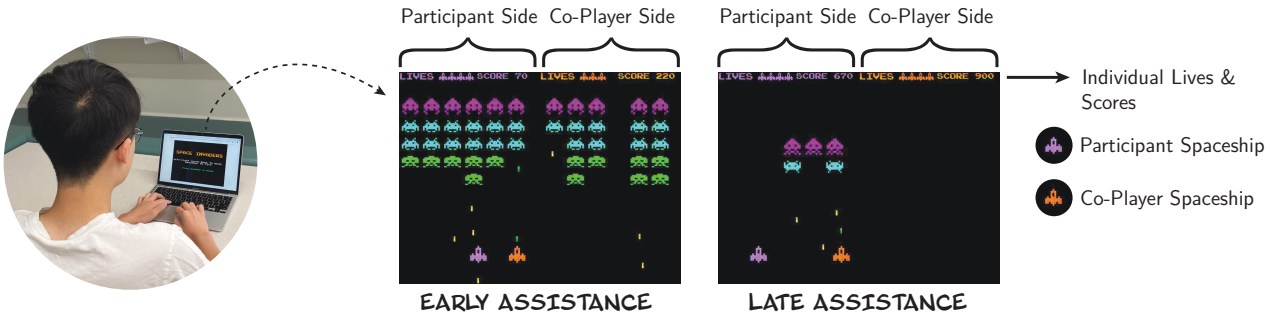
# Perceptions of the Helpfulness of Unexpected Agent Assistance

Kate Candon
Zoe Hsu
kate.candon@yale.edu
zoe.hsu@yale.edu
Yale University
New Haven, CT, USA

Yoony Kim
Jesse Chen
yoony.kim@yale.edu
jesse.b.chen@yale.edu
Yale University
New Haven, CT, USA

Nathan Tsoi
Marynel Vázquez
nathan.tsoi@yale.edu
marynel.vazquez@yale.edu
Yale University
New Haven, CT, USA




a) Participants played a multi-player Space Invaders game with one of three co-player identities.

b) Participants experienced 2 assistive behaviors by the co-player: it helped destroy participant enemies on the left side of the screen early in the game, or helped late after destroying its own enemies.

**Figure 1: The study investigated how participants perceived an agent that, unexpectedly, acted prosocially in a Space Invaders game. We controlled for the agent's assistive behavior (Early vs. Late Assistance) and its identity (Human vs. Computer vs. AI).**

## ABSTRACT

Much prior work on creating social agents that assist users relies on preconceived assumptions of what it means to be helpful. For example, it is common to assume that a helpful agent just assists with achieving a user's objective. However, as assistive agents become more widespread, human-agent interactions may be more ad-hoc, providing opportunities for unexpected agent assistance. How would this affect human notions of an agent's helpfulness? To investigate this question, we conducted an exploratory study (N=186) where participants interacted with agents displaying unexpected, assistive behaviors in a Space Invaders game and we studied factors that may influence perceived helpfulness in these interactions. Our results challenge the idea that human perceptions of the helpfulness of unexpected agent assistance can be derived from a universal, objective definition of help. Also, humans will reciprocate unexpected assistance, but might not always consider that they are in fact helping an agent. Based on our findings, we recommend considering personalization and adaptation when designing future assistive behaviors for prosocial agents that may try to help users in unexpected situations.


*HAI '22, December 5–8, 2022, Christchurch, New Zealand*

ACM ISBN 978-1-4503-9323-2/22/12.
https://doi.org/10.1145/3527188.3561915

## CCS CONCEPTS

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

## KEYWORDS

human-agent interaction, prosocial behavior, assistive agents

**ACM Reference Format:**
Kate Candon, Zoe Hsu, Yoony Kim, Jesse Chen, Nathan Tsoi, and Marynel Vázquez. 2022. Perceptions of the Helpfulness of Unexpected Agent Assistance. In *Proceedings of the 10th International Conference on Human-Agent Interaction (HAI '22), December 5–8, 2022, Christchurch, New Zealand.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3527188.3561915

## 1 INTRODUCTION

The notion of autonomous agents that help users is becoming more and more prevalent in every day life. Conversational assistants can help users find information on the web [43, 66], tools powered by Artificial Intelligence (AI) can support pathologists in making diagnoses [23], computer-mediated peer support can combat isolation in home care workers [52], and digital intervention systems can help build healthy eating habits [17] or treat compulsive gaming [68]. Across these scenarios, it is common to assume that the agent's only goal is to support the human and to define helping behaviors for autonomous agents in terms of application-specific user needs.

As assistive agents become more widespread, opportunities arise for human-agent interactions to be more ad-hoc with less defined dynamics between parties. For example, agents may begin acting individually in an environment and then find reasons to engage in cooperative activities with unfamiliar human teammates [57].

Agents may also find opportunities to assist others despite some personal cost, i.e., to engage in prosocial behavior [11, 13]. How would people interpret unexpected prosocial actions from an agent when they are not primed for assistance?

Inspired by prior work that investigates human-agent interactions via social games [3, 29, 38], we conducted an exploratory online study to understand how people perceived the helpfulness of an agent in a multi-player Space Invaders game, as shown in Fig. 1. In this game, players moved their spaceship and shot bullets across the game screen to destroy incoming enemies. An autonomous agent provided unexpected prosocial assistance by destroying enemies on the participant's side of the screen (see Fig. 1(b)). This resulted in more points for the participant in the game.

When the study began, there was no requirement for cooperation among players, nor expectation for assistance from the other agent in Space Invaders. That is, participants were not expecting the co-player to help them destroy enemies on the left side of the game screen. We purposefully did not incentivize teaming (or competition) with the autonomous agent, and intentionally chose a game with ambiguity around how players should behave in this regard, because this ambiguity is realistic in many real collaboration scenarios. For example, consider a scenario where you are watching someone cook, and you notice that the garlic in a pan is starting to burn. If you remove the pan from the heat, would they think you were helping or feel like you were interfering with their cooking?

We suspected that Space Invaders would allow us to observe varying human perceptions of agent helpfulness based on the type of assistive behavior that the agent provided during interactions, the order in which humans experienced these behaviors, and how the agent was described to the participants. In particular, we designed two types of helping behaviors for the agent, which differed based on the timing of the assistance (Fig. 1(b)). We also introduced the agent to the participant as either controlled by another human, by artificial intelligence, or by a computer.

By studying human-agent interactions in Space Invaders, we uncovered valuable insights for the development of future social agents. First, our findings challenge the conventional notion that helpful agents are those that simply achieve the human's objective. Instead, our work suggests that it is crucial for autonomous agents to reason about how their assistive actions are perceived because human notions of helpfulness are nonuniversal. What the designers of autonomous agents think is helpful behavior may not actually be perceived as helpful in practice.

## 2 RELATED WORK

**Games in Human-Agent Interaction:** Previous studies have used multi-player games to study human-agent interactions because games are easy to adapt to specific research needs and can reflect important aspects of real-world interactions. Games motivate users to engage and provide freedom to explore novel interaction methods that would otherwise not be feasible to implement at scale [62].

Prior work in Human-Agent Interaction (HAI) has typically considered interactions involving turn-taking or survival games that explicitly encourage cooperation [2, 25, 64] or competition [12, 50]. Less common are scenarios where neither setting is pre-established. One important exception is the work by Large et al. [38], which our study directly builds from. Their work studied how humans perceived and reacted to a cooperative and an uncooperative agent in a two-player version of Space Invaders. The cooperative agent helped a human destroy their enemies, and was found to be more helpful than the uncooperative agent, which only destroyed its own enemies. In the present work, we used a similar game to the one used by Large et al. [38] and did not encourage cooperation nor competition among players. Different from the previous work, we compared two types of assistive behaviors for an agent, studied the effects of how an agent was introduced, and further analyzed the nuances of perceived helpfulness.

**Agent Identity:** Past work has investigated how the identity of an agent may affect human perception of the agent and interactions with it. For example, Li et al. [41] explored how the identity of a lecturer affected video instruction and found that attitudes were more positive toward human lecturers than toward robots. In AI-mediated communication, Jakesch et al. [30] observed a "replicant effect" on how much people trusted hosts in the Airbnb platform: only when the participants thought that they saw a mixed set of host profiles written by AI and humans, did they mistrust hosts whose profiles were labeled as or suspected to be written by AI.

Closer to our work, Ashktorab et al. [3] investigated human-agent interactions with different agent identities in a cooperative word association game. Their results suggested that participants found the player labeled as human to be more likable and helpful than the player labeled as AI, but the identity had no impact on the outcome of the game. Additionally, even when bots do have superior abilities (e.g., negotiating), this advantage can be nullified by biases humans hold when participants are informed they are interacting with a bot [29]. In contrast to the discrete nature of the aforementioned games, our experimental task, Space Invaders, requires more continuous decision-making and has a fast pace.

**Agent Helpfulness:** Past research has extensively studied how interactive agents can be helpful by supporting users (e.g., [36, 53, 67, 68]) and facilitating better interactions within groups of humans (e.g., [15, 16, 24, 45, 56]). While past supportive technologies may be autonomous and responsive to specific events, it is common to create social agents using preconceived notions of what a "helpful" action will be. For example, Duan et al. [16] studied how an autonomous agent could improve collaborative interactions between native and non-native speakers. In this work, the agent was pre-programmed with a preconceived notion of what it meant to try to help during the interaction: asking for clarification when rare words were used in the conversation, subject to frequency constraints. However, one could imagine situations where this behavior might be undesirable. Asking for clarification could break the flow of the conversation if no one was actually confused.

Prior work highlights the importance of different factors that can affect human perception of assistance. One important factor is the timing of assistive actions [10, 31, 32], which motivated the two helping behaviors that we designed for an agent in Space Invaders. Additionally, Rudman and Zajicek [54] found that when balancing helpfulness and annoyingness, autonomous agents need to consider human feelings about interactions, ensuring that actions are not only objectively useful, but also perceived to be useful. Similarly,

Kim et al. [34] found that simply presenting help does not necessarily correspond to perceptions traditionally tied to assistance, such as agreeableness. Motivated by these works, our study investigated human perceptions of unexpected agent assistance in Space Invaders to better understand:

*RQ 1a: When help is unexpected, is the notion of helpfulness influenced by factors such as agent identity or the timing of assistance?*

*RQ 1b: How do other agent attributes relate to perceived helpfulness of unexpected assistance?* Examples of agent attributes include perceived intelligence or annoyingness.

**Prosocial Behavior:** This work is in part motivated by a longstanding interest in prosocial behavior in social psychology [46, 58] and behavioral economics [4]. Prior work has generally focused on how and when people help [47]. However, why people help is more complicated because the motivations driving individuals are varied and complex [33].

The question of how to create computational agents with social cognition abilities that are capable of rendering prosocial actions has inspired significant work within AI [8, 48]. One approach to creating helping agents is to implement a *social goal adoption* mechanism [8], where the internal goals of an agent change to be those of the user it is trying to help. This framing inspired the assistive behaviors that we chose for an agent in Space Invaders (Fig. 1(b)).

Research in cooperative and prosocial AI technologies has been diverse, covering coordination aspects [19, 37, 51, 60], teaming [44, 59], goal inference [65] or reward inference [69], and the development of multi-agent systems that cooperate [35, 39, 40]. Prior work challenges a utilitarian view of human decision-making and aims to better understand, predict, and promote prosocial behavior among humans through artificial agents [48]. This line of work motivated us to study human-agent interactions in a setting where assistance may emerge, but is not necessarily expected nor required.

There are gaps in current research to understand how effectively robots and virtual agents are able to promote prosocial behavior [47]. Because prosocial behavior has been tied to reciprocity [63], we decided to also investigate:

*RQ 2: Do assistive behaviors or agent identity influence reciprocity of unexpected assistive actions?*

# 3 METHOD

We conducted an exploratory online study to investigate the research questions outlined in the prior Section. In the study, the participants completed a web survey through which they interacted with an agent in a two-player version of Space Invaders (Fig. 1). Players tried to destroy enemies before they reached the bottom of the screen, or before running out of lives. Players lost a life if they collided with an enemy or were hit by a bullet from one of the enemies. The participants received points for enemies destroyed on the left half of the screen whereas the autonomous agent – referred to as the *co-player* hereafter – received points for enemies destroyed on the right half of the screen (as in Fig. 1(b)). The co-player was controlled by the same algorithm for all participants, but the participants were told that the co-player was controlled by a human, a computer, or by Artificial Intelligence (AI), as later explained in Sec. 3.3. The study was approved by our local Institutional Review Board and refined via pilots.

## 3.1 Participants

We recruited 360 participants for the study through Prolific [49].[1] Participants' recruitment criteria required them to be 18 years of age or older, be fluent in English, reside in the United States, and have normal or corrected-to-normal vision.

Out of the 360 participants, the study had a final sample size of 186 participants because 174 participants were excluded for several reasons. First, three participants asked to withdraw from the study after playing Space Invaders and being debriefed about the true identity of the agent. Second, the game logs for 141 participants indicated that in at least one round of Space Invaders, the co-player did not successfully destroy any of the enemies for which the participant received points due to slow Internet speed. Therefore, these participants did not experience the intended helping behavior by the co-player. Third, 22 participants were excluded due to data collection issues. Lastly, eight participants were excluded because they indicated in free response questions that they did not believe the co-player was another person in the human Identity condition, which went against our manipulation. More details about the exclusion criteria are provided in Sec. A of the Supplementary Material.

Of the 186 final participants, 78% identified as female, 20% identified as male, 1% identified as nonbinary, and 1% preferred not to say. The participants' ages ranged from 18 to 66 years old, with an average age of 26.92 years ($SD$ = 9.43). At the beginning of the study, the participants completed a demographics survey. In this survey, they indicated using computers often: 94% of participants used a computer daily, 6% used a computer 4-6 times a week, and <1% used a computer 2-3 times a week. A little more than half of the participants (53%) played video games at least once a week. Specifically in regards to Space Invaders, 46% of the participants reported that they had played the game before, 46% reported they had not played it, and 8% were unsure. Details on demographics by condition are provided in Sec. B of the Supplementary Material.

## 3.2 Space Invaders Game

The Space Invaders game can be seen in Fig. 1. The participant controlled a purple spaceship using the arrow keys and space bar on their keyboard. Their spaceship started on the left side of the game screen and it could shoot bullets upwards to destroy incoming enemies. The participant's co-player was an orange spaceship, which started on the right side of the game and could shoot upwards as well. The participant and co-player could move left and right within the full bounds of the game screen. The participant and co-player were assigned points individually for enemies destroyed on the side of the screen on which they originally started. Each player started with four lives, and lost a life when hit by an enemy or a bullet from the enemies. As the game progressed, enemies moved downwards, closer to the player until they were destroyed or reached the bottom of the game screen. The game ended when all enemies were destroyed, both players lost all their lives, or an enemy reached the bottom of the game screen. Unbeknownst to the participant, the co-player adjusted its shooting speed to match

[1]The number of participants recruited for the experiment was guided by a power analysis [18] which suggested a sample size of 158 participants. We recruited more participants than our power analysis indicated because we were collecting data remotely via an online survey and knew that varying Internet speeds, which could lead to other technical issues with Space Invaders, would be prevalent based on pilots.

the participant's shooting speed. This was important to make the assistive behaviors more consistent across participants.

## 3.3 Study Design

We designed the study considering two main independent variables:

– Co-player *Identity* (3 levels). The participant was told that the co-player was controlled by another **human**, by a **computer**, or by Artificial Intelligence (**AI**). The co-player identity was highlighted in the game instructions. Due to questions about general AI literacy (e.g., [42]), we included both the computer and AI condition to see if there were measurable differences in perceptions.

– Co-player assistive *Behavior* (2 levels). Based on research on the importance of timing in human-robot collaboration (e.g., [10, 31, 32]), we focused on timing as the differentiating factor between two helping behaviors. Specifically, we changed when the co-player tried to help the participant by destroying enemies on the left side of the gamescreen. In one game, the co-player exhibited an **early-assistance** behavior. It went to the left side of the screen twice to help destroy the participant's enemies before it finished destroying its own enemies on the right side of the screen. The co-player first moved to the participant's side of the screen once the co-player had destroyed 25% of its own enemies and stayed until half of the enemies on the participant's side were destroyed. The second visit was prompted when the participant had destroyed 70% of its own enemies, and the co-player moved back to its own side once all of the participant's enemies were destroyed. In the other game, the co-player exhibited a **late-assistance** behavior. Under this behavior, the co-player helped destroy the participant's enemies on the left side only after all of its own enemies on the right side were destroyed. Fig. 1(b) shows example screenshots of the two helping behaviors. In both cases, helping to destroy enemies on the left side of the gamescreen was considered prosocial assistance because the co-player scored points for the participant with no benefit to itself. The assistance was also unexpected because we did not prime participants for cooperation.

We used a 3 (Identity, between) x 2 (Behavior, within) mixed-design for the user study. That is, each participant experienced one co-player identity, but played two rounds of the game, one with each assistive behavior. The participants were not told which behavior they experienced in which game. It is worth noting that we counter-balanced the order in which the participants experienced the assistive behaviors because we suspected that this order could influence interactions and human perceptions of the co-player.

## 3.4 Procedure

The experiment was conducted via an online Qualtrics survey and included two games of Space Invaders. Upon the beginning of the survey, the participant consented to participate in the study.

Next, the survey asked for the participant's demographic information, as discussed in Sec. 3.1. The survey also gathered personality data via the Revised Competitiveness Index [27] and the Ten Item Personality Measure (TIPI) [21]. Then, the survey introduced the Space Invaders game with a combination of text explanations and visual instructions (see pages 11 and 12 of the supplementary survey text for examples). Importantly, the text conveyed the co-player's identity to the participant (either human, computer or AI). We purposefully did not tell the participant that they would be helped by the co-player or that they could help it during the game because we did not want to prime the participant for cooperation.

The participant experienced two games of Space Invaders, each followed by a post-game survey about their experience and their perceptions of the co-player. One game showed the early-assistance behavior and the other showed the late-assistance behavior, but the participants were not informed of the order in which they experienced the two behaviors.

At the end of the study, the participants were asked a final set of questions about the differences between the games. Also, the participants who were told that the co-player was human were debriefed by telling them that the co-player was automatically controlled by an algorithm. Due to the deception, the participants in the human condition were reminded that they had the option to withdraw from the study. Finally, the survey presented an optional Berkeley Expressivity Questionnaire (BEQ) [22] and the participants had an open response question to report any bugs or other comments to the researchers. Participants were paid $3.60 to participate in the study, whether or not they chose to withdraw at the end of the survey. The study typically took about 18 minutes to complete.

## 3.5 Dependent Measures

The study relied on a combination of qualitative and quantitative measures to analyze the participants' perceptions of the interaction and their reactions to the assistive behaviors from the co-player during the game. We chose to investigate the perceptions of the co-player in an exploratory manner. During the study, we measured factors related to the perception of the co-player (such as its competence) in addition to factors related to perception of the Space Invaders game (such as its perceived level of difficulty). Our aim was to advance our understanding of the complexity of human perceptions of helping behaviors during a task that involves continuous decision making.

*3.5.1 Survey Questions.* Survey questions included questions administered right after a given Space Invaders game, as well as the final questions about both of the games were played:

- **Perception of Co-player:** After each game of Space Invaders, the participants were asked to rate their agreement with statements about the co-player on a scale from 1 (strongly disagree) to 7 (strongly agree). The five statements, from Large et al. [38], were "the co-player was helpful", "the co-player was proficient", "the co-player was intelligent", "the co-player was annoying", and "I liked the behavior of the co-player in the game". In addition, the participants were asked to evaluate the level of warmth, competence, and discomfort of the co-player using the 18 attributes from the Robotic Social Attributes Scale (RoSAS) [7]. These ratings were obtained on a scale from 1 (not at all) to 7 (very much so), and had high reliability for all 3 subscales. In particular, Warmth had a Cronbach's $\alpha = .88$, Competence had $\alpha = .87$, and Discomfort had $\alpha = .79$. Lastly, the participants were also asked if anything about the behavior of the co-player seemed unusual.

- **Help:** After each game of Space Invaders, the participants were asked if they helped the co-player. They could respond "Yes", "No", or "Not sure" and explain their rationale.
- **Game Experience:** After each game of Space Invaders, the participants were asked to rate their agreement with four statements about the game on a scale from 1 (strongly disagree) to 7 (strongly agree). The statements, again from Large et al. [38], were: "I enjoyed the game", "the game was difficult", "the game was boring", and "I would play this game for fun".

*3.5.2 Participant Actions in the Game.* We analyzed events of the games using game logs. These game logs contained information about the state of the game, participant actions, and co-player actions for each rendered frame of Space Invaders. In particular, we extracted and analyzed: the number of participant enemies destroyed by the co-player, the number of co-player enemies destroyed by the participant, and the participant and co-player's final scores and number of lives remaining.

## 3.6 Implementation Details

The Space Invaders game was implemented using a combination of browser-based client technologies along with a Python server. For the client, we used the Phaser game framework[2], which is built on HTML5 and designed to run in a web browser. The Phaser-based client was responsible for rendering the game and accepting user input. Communication between the client and the server was facilitated by a bi-directional, real-time connection using the WebSocket Protocol.[3] Our server, based on the Tornado web framework, received user input and game state to determine the next action for the co-player. We hosted our Space Invaders game on a computer with 8GB of RAM and 4 cores clocked at 2.3GHz.

# 4 RESULTS

This section describes our analyses of post-game survey measures, final survey questions, and game logs to understand participant's notions of helpfulness of the co-player. Perceived helpfulness corresponded to participants' agreement with "The co-player was helpful", as described in Sec. 3.5. Unless otherwise noted, we used Restricted Maximum Likelihood (REML) analyses [61] via JMP Pro [28] to statistically examine survey data. In these analyses, co-player Behavior (early-assistance or late-assistance), Order (early-first or late-first), and co-player Identity (human, AI, or computer) were considered as main effects, and Participant ID was a random effect. We conducted post-hoc Tukey Honestly Significant Difference (HSD) or Student's t-tests when appropriate.

## 4.1 Perceptions of Helpfulness by Agent Behavior and Identity

We first investigated the effects of Behavior, Order, and Identity on participant's perceived helpfulness for the co-player. We also used correlation analyses to investigate the relationship between helpfulness ratings and effective co-player actions to destroy enemies for which the participant received points in the game.

[2]https://phaser.io
[3]https://developer.mozilla.org/en-US/docs/Web/API/WebSocket

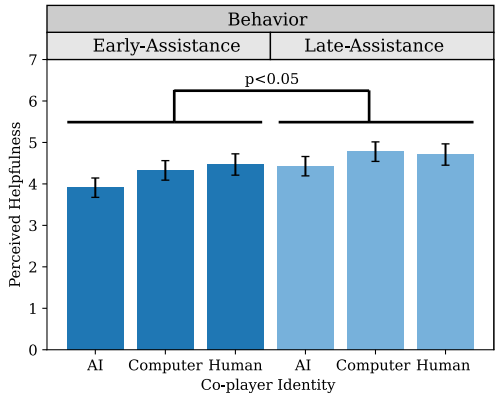



**Figure 2: Perceived helpfulness of the co-player by Behavior and Identity. Perceived helpfulness corresponds to agreement with "The co-player was helpful" on a 7-point scale.**

*4.1.1 Effects of Behavior, Order, and Identity on Co-Player Helpfulness.* There was a significant difference in how helpful participants rated each of the two helping Behaviors, $F(1, 182) = 5.70, p = .018$, as shown in Fig. 2. The participants rated the late-assistance co-player ($M = 4.64, SE = .14$) as significantly more helpful than the early-assistance co-player ($M = 4.23, SE = .14$). There was no significant effect from the Order in which participants experienced the behaviors, $p = .59$. We similarly found no significant effect of co-player Identity on how helpful the participant rated the co-player's behavior, $p = .23$.

Interestingly, there was a significant interaction between Behavior and Order on co-player helpfulness, $F(1, 182) = 12.80, p = 0.0004$. A Tukey's HSD post-hoc test revealed that for the early-first participants, the late-assistance co-player ($M = 5.00, SE = .20$) was perceived as more helpful than the early-assistance co-player ($M = 3.99, SE = .20$), $p = .0003$. No other pair-wise significant differences in helpfulness were observed.

*4.1.2 Perceived Helpfulness and Objectively Helpful Actions.* Because of the design of Space Invaders, one is naturally tempted to associate co-player helpfulness to the participant's objective: destroying enemies on their side of the screen. In that case, one would expect a positive correlation between the number of enemies on the participant's side that the co-player destroyed and how helpful the participant perceived the agent to be. However, over both games for all participants, there was not a significant correlation between the number of participant's enemies that the co-player helped destroy, as recorded by the game logs, and how strongly the participant agreed that the co-player was helpful ($r(372) = -.06, p = .29$). Similarly, we found no significant correlation within co-player Behaviors (early-assistance: $r(186) = .03, p = .69$; late-assistance: $r(186) = -.04, p = .63$), Order (early-first: $r(178) = -.12, p = .10$; late-first: $r(194) = .02, p = .78$), and co-player Identity groups (AI: $r(132) = -.16, p = .07$; computer: $r(130) = -.01, p = .90$; human: $r(110) = -.02, p = .85$). Lastly, we explored whether there was a correlation between final participant score and perceived helpfulness across all participants, but found no significant correlation in this case either ($r(372) = .02, p = .77$).

## 4.2 Helpfulness and Other Agent Attributes

We investigated whether helpfulness was correlated with other relevant co-player attributes measured via survey questions. Across all participants, we found a strong correlation between the participant's perception of the co-player's helpfulness and how much they liked the co-player's behavior, ($r(372) = .63, p < .0001$). The next strongest correlations with helpfulness were a negative correlation with annoyance ($r(372) = -.47, p < .0001$) and a positive correlation with competence ($r(372) = .44, p < .0001$). More details about correlations between human perceptions of other agent attributes are discussed in Sec. C.1 of the Supplementary Material.

We also explored if correlations between helpfulness and other co-player attributes varied across Identity, Behavior, and Order. Results for Identity are shown in Table 1. While the strength of relative correlations is fairly consistent across co-player identities, there are some differences. Most notably, the correlation between helpfulness and discomfort is significant for participants in the human and computer groups, but not for the AI group. Also, the correlation between helpfulness and proficiency is significant for the human and AI groups, but not for the computer group. There were no changes in significance of correlations for Behavior or Order (see Sec. C.2 of the Supplementary Materials).

## 4.3 Participant Reciprocity

We investigated whether participants reciprocated help from the co-player by destroying enemies on the right side of the screen.

*4.3.1 Effects of Behavior, Order, and Identity on Reciprocity.* An REML analysis on the number of the co-player's enemies that participants destroyed in Space Invaders indicated a significant effect of Behavior on this number, $F(1, 182) = 138, p < .0001$. In particular, the participants destroyed significantly more of the co-player's enemies when experiencing the early-assistance co-player ($M = 3.13, SE = .18$) than when experiencing the late-assistance co-player ($M = .13, SE = .18$). Neither co-player Identity ($p = .053$) nor Order ($p = .36$) resulted in significant effects.

*4.3.2 Reciprocity with Early-Assistance Co-player.* With the early-assistance co-player, 129 participants (69%) of participants destroyed at least one enemy for which the co-player received points. Notably, a Chi-square test of independence showed that the likelihood a participant destroyed at least one of the early-assistance co-player's enemies differed by Order, $\chi^2(1, N = 186) = 4.59, p = .038$, but not by Identity, $\chi^2(2, N = 186) = 4.362, p = .112$. A Fisher's Exact Test showed that the likelihood a participant destroyed at least one of the co-player's enemies was higher when participants played the late-assistance co-player first than when they played the early-assistance co-player first, $p = 0.024$. When they played the early-assistance co-player in their second game, 76% of participants destroyed at least one of the co-player's enemies, compared to 62% of participants who played the early-assistance co-player in their first game. The 129 participants who destroyed at least one of the early-assistance co-player's enemies destroyed an average of 4.55 co-player enemies ($SE = .29$, maximum value = 16). Notably, all 129 participants did so only after the early-assistance co-player had come over to the participant's side to help. In response to *"Did you help the co-player?"* after experiencing the early-assistance behavior, 64% (83) of the participants that helped selected "yes", 27% (35) selected "no", and 9% (11) selected "not sure". This highlights that not all participants considered destroying co-player enemies as helping the co-player, so what is considered helpful can be ambiguous.

**Table 1: Correlations between perceived co-player helpfulness ratings and other co-player attributes (one per row).**

| | | Identity | | |
|---|---|---|---|---|
| Attribute | All | AI | Computer | Human |
| Proficiency | 0.3199 **** | 0.119 NS | 0.4801 **** | 0.3374 *** |
| Intelligence | 0.3382 **** | 0.2802 ** | 0.4767 **** | 0.2565 * |
| Annoyingness | -0.4661 **** | -0.4406 **** | -0.383 **** | -0.5848 **** |
| Liked Behavior | 0.6271 **** | 0.5813 **** | 0.607 **** | 0.6975 **** |
| Warmth | 0.3585 **** | 0.2576 ** | 0.4385 **** | 0.3739 **** |
| Competence | 0.4416 **** | 0.3731 **** | 0.546 **** | 0.3856 **** |
| Discomfort | -0.2621 **** | -0.3214 *** | -0.1661 NS | -0.2884 ** |

* p < .05. ** p < .005. *** p < .0005. **** p < .0001 NS not significant

For the 129 participants who did reciprocate assistance, we analyzed their responses to "Please explain your answer about if you helped the co-player." We first separated responses by self-reported response to "Did you help the co-player?", and then clustered responses into themes via an affinity diagram. Number of responses per theme, by co-player Identity and Order, are shown in Table 2. Of the participants who responded that they did help the co-player, the most common rationale was because they had destroyed all of the enemies on their own side. For example, P0414Z wrote *"I went to their side to help them out after finishing off my enemies."* Reciprocity was the third most common rationale for providing help. P0226Z wrote *"The co-player came over and helped me before he was even finished with his side. It felt only right to go over and help him finish his."* Within participants who answered that they did not help the co-player, the two most meaningful themes were that they were too focused on their own gameplay and that they felt they were competing with the co-player. P0262Z responded: *"It was competition so why should I help them"*, and P0263Z responded *"I was focused on myself not the co-player."* Notably, five of the eleven participants who responded they were not sure if they helped the co-player suggested that they were not sure what was considered help in this setting, e.g., *"I'm not sure what is considered helping the other player."* (P0374Z).

*4.3.3 Reciprocity with Late-Assistance Co-player.* In the games with the late-assistance co-player, only ten (5%) of the participants destroyed at least one of the co-players enemies ($M = 2.60, SE = .76$, maximum value = 9). Notably, eight of the ten participants were in the Early-First Order group, so they had received help from the agent in the previous game. The two participants in the Late-First Order group did not have the option to "reciprocate" help because they would not have seen the co-player come help until there were no co-player enemies left. Of these two participants, one responded that they did help the co-player because they *"thought of us as a team"*. The other participant answered that they did not help the agent, and in fact destroyed only one of the co-player's enemies in the left-most column of the right half of enemies (i.e., towards the middle of the screen) after losing a life. When the life was lost, the player respawned in the middle of the screen, so they may not have realized that they had destroyed one of the co-player's enemies.

**Table 2: Explanations from the 129 participants who destroyed at least one co-player enemy about why they did or did not help the early-assistance co-player. Responses were separated by response to "Did you help the co-player?". Counts were broken down by Identity: AI (AI), Computer (C), Human (H); and by Order: Early-First (EF), Late-First (LF).**

| Did you help the co-player? | Identity | | | Order | | |
|---|---|---|---|---|---|---|
| **Theme** | **AI** | **C** | **H** | **EF** | **LF** | **Tot** |
| ***Yes*** | ***35*** | ***26*** | ***22*** | ***27*** | ***56*** | ***83*** |
| Done with own side | 12 | 8 | 5 | 9 | 16 | 25 |
| No reason was provided | 7 | 7 | 9 | 9 | 14 | 23 |
| Reciprocity | 7 | 3 | 2 | 3 | 9 | 12 |
| Team | 4 | 3 | 2 | 4 | 5 | 9 |
| Make game end faster | 3 | 1 | 0 | 1 | 3 | 4 |
| Swap sides | 0 | 2 | 1 | 0 | 3 | 3 |
| Better than co-player | 0 | 0 | 2 | 0 | 2 | 2 |
| New awareness of ability | 0 | 1 | 1 | 0 | 2 | 2 |
| Other | 2 | 1 | 0 | 1 | 2 | 3 |
| ***No*** | ***15*** | ***10*** | ***10*** | ***20*** | ***15*** | ***35*** |
| No reason was provided | 7 | 3 | 5 | 9 | 6 | 15 |
| Focused on own gameplay | 3 | 4 | 1 | 4 | 4 | 8 |
| Competition | 5 | 0 | 2 | 6 | 1 | 7 |
| Co-player was adversarial | 0 | 2 | 1 | 0 | 3 | 3 |
| Other | 0 | 1 | 1 | 1 | 1 | 2 |
| ***Not sure*** | ***2*** | ***5*** | ***4*** | ***8*** | ***3*** | ***11*** |
| Unsure of helpful definition | 1 | 2 | 2 | 3 | 2 | 5 |
| Not sure if helped | 1 | 1 | 1 | 3 | 0 | 3 |
| Other | 0 | 2 | 1 | 2 | 1 | 3 |

## 4.4 Other Findings

Due to the exploratory nature of our study, we expanded our investigation beyond co-player helpfulness by analyzing post-game survey responses in relation to the Space Invaders game and perceived co-player attributes. Neither co-player Identity, Behavior or Order resulted in significant differences with respect to the four game experience statements that were part of our post-game survey questions (i.e., enjoy, difficult, boring, and would play for fun). Overall, participants enjoyed the game ($M = 5.23, SE = .07$) and would play the game for fun ($M = 4.68, SE = .10$). They did not find the game difficult ($M = 3.45, SE = .09$) or boring ($M = 2.65, SE = .08$).

Identity had a significant effect on how warmly the co-player was perceived, $F(2, 180) = 7.14, p = .001$. The participants in the human Identity group ($M = 3.01, SE = .16$) perceived the co-player as significantly warmer than those in the computer ($M = 2.30, SE = .15$) and AI ($M = 2.27, SE = .15$) groups. We found no other significant effects of Identity on other perceptions of the co-player.

While REML analyses indicated only one significant difference in perceptions of the co-player based on co-player Identity, we found several differences based on co-player Behavior. Table 3 shows the results for the co-player attributes that we examined, including helpfulness as discussed in Sec. 4.1, by co-player Behaviors across all participants. On average, the participants rated the late-assistance co-player as significantly more helpful, proficient, intelligent, and competent than the early-assistance co-player. The late-assistance co-player was also rated as significantly less annoying and less discomforting than the early-assistance co-player. Based on responses to *"I liked the behavior of the co-player in the game"*, the participants liked the late-assistance behavior significantly more than the early-assistance behavior, on average. Warmth was the one dimension we evaluated where there was not a significant effect from Behavior.

**Table 3: Means (M), Standard Error (SE) and F-statistic (F) from REML analyses on the effect of co-player Behavior on agent attributes. The measures correspond to participant agreement with statements about the co-player on a 7-point scale. The results consider all participants.**

| Measures | Early-Assistance | | Late-Assistance | | F(1,190) |
|---|---|---|---|---|---|
| | M | SE | M | SE | |
| Helpfulness | 4.23 | 0.14 | 4.64 | 0.14 | 5.70* |
| Proficiency | 5.16 | 0.10 | 6.04 | 0.10 | 53.26**** |
| Intelligence | 4.54 | 0.11 | 5.16 | 0.11 | 24.92**** |
| Annoyance | 3.61 | 0.13 | 2.77 | 0.13 | 26.12**** |
| Liked behavior | 4.12 | 0.12 | 4.78 | 0.12 | 21.73**** |
| Warmth | 2.49 | 0.09 | 2.56 | 0.09 | 1.19 |
| Competence | 4.48 | 0.10 | 5.13 | 0.10 | 48.39**** |
| Discomfort | 2.38 | 0.08 | 2.05 | 0.08 | 20.56**** |

*$p < 0.05$. ****$p < 0.0001$.

There were no significant differences across agent attributes with respect to the Order in which participants experienced the co-player behaviors. However, similarly to how the interaction between the co-player Behavior and Order was significant for helpfulness, this interaction also had a significant effect on perceived discomfort, $F(1, 182) = 9.22, p = 0.003$. A Tukey's HSD test revealed that for early-first participants, they perceived the early-assistance co-player ($M = 2.48, SE = .11$) as significantly more discomforting than the late-assistance co-player ($M = 1.96, SE = .11$), $p < .0001$. There was no significant difference in discomfort when the late-assistance behavior was experienced first.

## 5 DISCUSSION

For *RQ1a*, we investigated whether the timing of assistance or agent identity influenced the perceived helpfulness of a co-player in Space Invaders. We found that participants rated the late-assistance co-player as more helpful than the early-assistance co-player. This was somewhat surprising because the late-assistance co-player only helped the participant score points by destroying enemies on the left side of the gamescreen once near the end of the game, compared to the early-assistance co-player who helped the participant score points by destroying enemies on the left side of the gamescreen on two earlier occasions during the game. However, it could be argued that the actions from the late-assistance behavior were more logical. After the co-player had finished destroying its enemies, it was free to go help the other player. Another explanation is that the late-assistance player reduced the stress of the participant by ensuring the enemies on its own side did not reach the bottom of

the screen to end the game. From this perspective, staying on the right side of the gamescreen could then be seen as helpful. However, we did not incentivize a specific goal in the participants, so a longer game was not necessarily better for all of them.

We also found that participants who experienced the early-assistance co-player first perceived the late-assistance co-player as more helpful and less discomforting than the early-assistance co-player. However, this was not true for the reverse order. This suggests that a human's prior experiences interacting with an agent may change their perceptions of its behaviors. Without expecting any assistance, the early-assistance behavior may have been comparatively more surprising than the late-assistance behavior.

REML analyses showed that co-player identity had a significant effect on how warmly participants perceived the co-player, but not on helpfulness as we had expected for *RQ1a*. Several participants remarked in our survey that they wished they could have texted with the co-player, so it is possible that more communication between players could have made the co-player identity more salient.

Regarding *RQ1b*, we found that helpfulness may be more personal and emotional than solely related to achieving objectives. Although we assumed that the reward system of Space Invaders would drive notions of helpfulness, we did not find a significant correlation between how many participant enemies the co-player destroyed and how helpful humans perceived this virtual agent to be. In addition, we found that the co-player's helpfulness was more strongly correlated with whether participants liked the agent's behavior and found it annoying than with its proficiency and intelligence. Annoyingness, in particular, is more personal, whereas proficiency is more closely tied to achieving a goal.

It was interesting to see that helpfulness was not significantly correlated with the same agent attributes across identities. For example, helpfulness was significantly correlated with proficiency for a computer or human agent, but not for an AI agent. Also, helpfulness had a significant negative correlation with discomfort for a human or AI agent, but the correlation was not significant for a computer agent. These different results could be because participants have biases about the proficiency of AI or do not consider discomfort from a computer in the same way as for the other identities. Future work should further investigate how identity may influence perceptions of prosocial agents.

Regarding *RQ2*, our results indicated that agent behavior had a significant effect on participant reciprocity of assistive actions, but identity did not. Participants destroyed more co-player's enemies when they experienced the early-assistance than the late-assistance behavior. While many participants indicated helping because they were done with their own enemies, 12 participants (14% of those that indicated helping) said they helped because they were reciprocating assistance. Also, nine participants (11%) said that they helped because they felt like a team, even though we never asked participants to collaborate with the co-player. This suggests that unexpected assistance could be a mechanism to motivate human cooperation [5] and, more generally, engineer prosociality [48].

Taken together, all of these results support an overarching idea: there is more to helpfulness than solely supporting another agent in achieving their goal. Even though the early-assistance co-player took more actions to help the participant score points, participants rated the late-assistance co-player as more helpful. While 69% of participants reciprocated help in the game with the early-assistance co-player, less than two-thirds of them reported that they believed they had in fact helped the co-player, suggesting the definition of help was ambiguous. Qualitative analyses also reinforced the idea that participants can have differing opinions of what it means to help, even in a simple environment such as Space Invaders.

A line of research that is worth discussing in relation to our work are efforts to develop robots for physical assistance, e.g., to help people move from one location to another [20] or manipulate the physical state of the world [6, 26]. While these robots are often designed with teleoperation interfaces such that users can directly control them, much recent work has contributed assistive teloperation algorithms where effective assistance is user-dependent and requires online adaptation [1, 14]. Our work suggests that this personalized view of assistance is important for autonomous agents more generally, especially when agents' assistance is unexpected by users – which is not the case with physically assistive devices.

Because there might not be a universal definition for help, it is important to enable computational agents that assist users with the ability to personalize their behavior as interactions unfold. Research in continual learning for HAI provides a pathway to creating such adaptive agents [9]. Also, recent work in value alignment provides mechanisms to learn behaviors that match human desires [55].

## 6 LIMITATIONS AND FUTURE WORK

Our work is limited in two ways that motivate interesting future research directions. First, although the notion of early or late assistance is relevant in many applications, our analyses are bound to the domain of Space Invaders. It would be interesting to investigate unexpected agent assistance in other interactive scenarios, (e.g., with robots). Second, our investigation is limited by our protocol. The study was conducted as an online survey and we had to disregard the data of 48% of participants, mostly due to technical challenges. Improving the robustness of systems for crowdsourcing human-agent interactions could facilitate future research in understanding how to design and implement assistive social agents.

## 7 CONCLUSION

We conducted an exploratory online survey to investigate what factors influenced how participants perceived and reacted to unexpected help from an interactive agent in a Space Invaders game. We found that even in Space Invaders – a continuously-updating but structured domain – participants' interpretations of the co-player's actions and impressions of the game were highly nuanced. Our results suggest there may not be universal truths when it comes to understanding whether an agent's assistance will be received in a positive or negative manner when this assistance is unexpected, or even what actions we can assume humans will interpret as helpful. Rather, it is important to understand the individual and to adjust to what is influencing their perception of the interaction.

## ACKNOWLEDGMENTS

Thank you to the National Science Foundation (IIS-2106690) for partially supporting this work. Z. Hsu, Y. Kim, and J. Chen were partially supported by the Yale STARS program, the Yale College Dean's Research Fellowship, and the Yale Summer Experience Award.

# REFERENCES


[1] Reuben M Aronson and Henny Admoni. 2020. Eye gaze for assistive manipulation. In *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*. 552–554.

[2] 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.* Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3411764.3445256

[3] 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.

[4] Roland Bénabou and Jean Tirole. 2006. Incentives and prosocial behavior. *American economic review* 96, 5 (2006), 1652–1678.

[5] Terence C Burnham and Brian Hare. 2007. Engineering human cooperation. *Human nature* 18, 2 (2007), 88–108.

[6] Maria E Cabrera, Tapomayukh Bhattacharjee, Kavi Dey, and Maya Cakmak. 2021. An exploration of accessible remote tele-operation for assistive mobile manipulators in the home. In *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*. IEEE, 1202–1209.

[7] 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.

[8] Cristiano Castelfranchi. 1998. Modelling social action for AI agents. *Artificial intelligence* 103, 1-2 (1998), 157–182.

[9] Nikhil Churamani, Sinan Kalkan, and Hatice Gunes. 2020. Continual Learning for Affective Robotics: Why, What and How?. In *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 425–431.

[10] F Cini, T Banfi, G Ciuti, L Craighero, and M Controzzi. 2021. The relevance of signal timing in human-robot collaborative manipulation. *Science Robotics* 6, 58 (2021), eabg1308.

[11] Filipa Correia, Samuel F Mascarenhas, Samuel Gomes, Patrícia Arriaga, Iolanda Leite, Rui Prada, Francisco S Melo, and Ana Paiva. 2019. Exploring prosociality in human-robot teams. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 143–151.

[12] Rahul R Divekar, Hui Su, Jeffrey O Kephart, Maira Gratti DeBayser, Melina Guerra, Xiangyang Mou, Matthew Peveler, and Lisha Chen. 2020. Humaine: Human multi-agent immersive negotiation competition. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–10.

[13] John F Dovidio, Jane Allyn Piliavin, David A Schroeder, and Louis A Penner. 2017. *The social psychology of prosocial behavior.* Psychology Press.

[14] Anca D Dragan, Siddhartha S Srinivasa, and Kenton CT Lee. 2013. Teleoperation with intelligent and customizable interfaces. *Journal of Human-Robot Interaction* 2, 2 (2013), 33–57.

[15] Wen Duan, Naomi Yamashita, and Susan R. Fussell. 2019. Increasing Native Speakers' Awareness of the Need to Slow Down in Multilingual Conversations Using a Real-Time Speech Speedometer. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 171 (nov 2019), 25 pages. https://doi.org/10.1145/3359273

[16] Wen Duan, Naomi Yamashita, Yoshinari Shirai, and Susan R Fussell. 2021. Bridging Fluency Disparity between Native and Nonnative Speakers in Multilingual Multiparty Collaboration Using a Clarification Agent. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–31.

[17] Daniel A Epstein, Felicia Cordeiro, James Fogarty, Gary Hsieh, and Sean A Munson. 2016. Crumbs: lightweight daily food challenges to promote engagement and mindfulness. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 5632–5644.

[18] Franz Faul, Edgar Erdfelder, Albert-Georg Lang, and Axel Buchner. 2007. G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior research methods* 39, 2 (2007), 175–191.

[19] Asaf Frieder, Raz Lin, and Sarit Kraus. 2012. Agent-human coordination with communication costs under uncertainty. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 26.

[20] Aditya Goil, Matthew Derry, and Brenna D Argall. 2013. Using machine learning to blend human and robot controls for assisted wheelchair navigation. In *2013 IEEE 13th International Conference on Rehabilitation Robotics (ICORR)*. IEEE, 1–6.

[21] Samuel D Gosling, Peter J Rentfrow, and William B Swann Jr. 2003. A very brief measure of the Big-Five personality domains. *Journal of Research in personality* 37, 6 (2003), 504–528.

[22] James J Gross, OP John, and J Richards. 1995. *Berkeley expressivity questionnaire.* Edwin Mellen Press Lewiston, NY.

[23] Hongyan Gu, Jingbin Huang, Lauren Hung, and Xiang'Anthony' Chen. 2021. Lessons learned from designing an AI-enabled diagnosis tool for pathologists. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–25.

[24] Carl Gutwin, Scott Bateman, Gaurav Arora, and Ashley Coveney. 2017. Looking Away and Catching Up: Dealing with Brief Attentional Disconnection in Synchronous Groupware. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (Portland, Oregon, USA) *(CSCW '17)*. Association for Computing Machinery, New York, NY, USA, 2221–2235. https://doi.org/10.1145/2998181.2998226

[25] Feyza Merve Hafizoglu and Sandip Sen. 2018. Reputation based trust in human-agent teamwork without explicit coordination. In *Proceedings of the 6th International Conference on Human-Agent Interaction*. 238–245.

[26] Laura V Herlant, Rachel M Holladay, and Siddhartha S Srinivasa. 2016. Assistive teleoperation of robot arms via automatic time-optimal mode switching. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 35–42.

[27] John Houston, Paul Harris, Sandra McIntire, and Dientje Francis. 2002. Revising the competitiveness index using factor analysis. *Psychological Reports* 90, 1 (2002), 31–34.

[28] SAS Institute. 2021. JMP Pro version 15.0.0.

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

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

[31] Lars Christian Jensen, Kerstin Fischer, Franziska Kirstein, Dadhichi Shukla, Özgur Erkennt, and Justus Piater. 2017. It gets worse before it gets better: Timing of instructions in close human-robot collaboration. In *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*. 145–146.

[32] Lars Christian Jensen, Kerstin Fischer, Stefan-Daniel Suvei, and Leon Bodenhagen. 2017. Timing of multimodal robot behaviors during human-robot collaboration. In *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 1061–1066.

[33] Joy Johnson, Marthaly Irizarry, Nhu Nguyen, and Peter Maloney. 2018. Part 1: Foundational theories of human motivation. (2018).

[34] Jieun Kim, Young June Sah, and Hayeon Song. 2021. Agreeableness of a Virtual Agent: Effects of Reciprocity and Need for Help. In *2021 15th International Conference on Ubiquitous Information Management and Communication (IMCOM)*. IEEE, 1–6.

[35] Max Kleiman-Weiner, Mark K Ho, Joseph L Austerweil, Michael L Littman, and Joshua B Tenenbaum. 2016. Coordinate to cooperate or compete: abstract goals and joint intentions in social interaction. In *CogSci*.

[36] Thomas Kosch, Markus Funk, Albrecht Schmidt, and Lewis L. Chuang. 2018. Identifying Cognitive Assistance with Mobile Electroencephalography: A Case Study with In-Situ Projections for Manual Assembly. 2, EICS, Article 11 (jun 2018), 20 pages. https://doi.org/10.1145/3229093

[37] Peter Krafft, Chris Baker, Alex Pentland, and Joshua Tenenbaum. 2016. Modeling human ad hoc coordination. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 30.

[38] Jamie Large, Graham Stodolski, and Marynel Vázquez. 2020. Studying Human-Agent Interactions in Space Invaders. In *Proceedings of the 8th International Conference on Human-Agent Interaction*. 245–247.

[39] Joel Z Leibo, Vinicius Zambaldi, Marc Lanctot, Janusz Marecki, and Thore Graepel. 2017. Multi-agent reinforcement learning in sequential social dilemmas. *arXiv preprint arXiv:1702.03037* (2017).

[40] Adam Lerer and Alexander Peysakhovich. 2017. Maintaining cooperation in complex social dilemmas using deep reinforcement learning. *arXiv preprint arXiv:1707.01068* (2017).

[41] 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.

[42] Duri Long and Brian Magerko. 2020. What is AI literacy? Competencies and design considerations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–16.

[43] Silvia B Lovato, Anne Marie Piper, and Ellen A Wartella. 2019. Hey Google, do unicorns exist? Conversational agents as a path to answers to children's questions. In *Proceedings of the 18th ACM International Conference on Interaction Design and Children*. 301–313.

[44] Amar R Marathe, Kristin E Schaefer, Arthur W Evans, and Jason S Metcalfe. 2018. Bidirectional communication for effective human-agent teaming. In *International Conference on Virtual, Augmented and Mixed Reality*. Springer, 338–350.

[45] Moira McGregor and John C. Tang. 2017. More to Meetings: Challenges in Using Speech-Based Technology to Support Meetings. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (Portland, Oregon, USA) *(CSCW '17)*. Association for Computing Machinery, New York, NY, USA, 2208–2220. https://doi.org/10.1145/2998181.2998335

[46] Leo Montada. 1991. *Altruism in social systems*. Lewiston, NY; Toronto: Hogrefe & Huber.

[47] Raquel Oliveira, Patrícia Arriaga, Fernando P Santos, Samuel Mascarenhas, and Ana Paiva. 2021. Towards prosocial design: A scoping review of the use of robots and virtual agents to trigger prosocial behaviour. *Computers in Human Behavior* 114 (2021), 106547.

[48] Ana Paiva, Fernando Santos, and Francisco Santos. 2018. Engineering pro-sociality with autonomous agents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

[49] Stefan Palan and Christian Schitter. 2018. Prolific. ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance* 17 (2018), 22–27.

[50] Andre Pereira, Rui Prada, and Ana Paiva. 2014. Improving social presence in human-agent interaction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1449–1458.

[51] Patrick M Pilarski, Andrew Butcher, Michael Johanson, Matthew M Botvinick, Andrew Bolt, and Adam SR Parker. 2019. Learned human-agent decision-making, communication and joint action in a virtual reality environment. *arXiv preprint arXiv:1905.02691* (2019).

[52] Anthony Poon, Vaidehi Hussain, Julia Loughman, Ariel C Avgar, Madeline Sterling, and Nicola Dell. 2021. Computer-Mediated Peer Support Needs of Home Care Workers: Emotional Labor & the Politics of Professionalism. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–32.

[53] André Rodrigues, André R.B. Santos, Kyle Montague, and Tiago Guerreiro. 2021. Promoting Self-Efficacy Through an Effective Human-Powered Nonvisual Smartphone Task Assistant. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 114 (apr 2021), 19 pages. https://doi.org/10.1145/3449188

[54] Paul Rudman and Mary Zajicek. 2006. Autonomous agent as helper-helpful or annoying?. In *2006 IEEE/WIC/ACM International Conference on Intelligent Agent Technology*. IEEE, 170–176.

[55] Stuart Russell. 2017. Provably beneficial artificial intelligence. The Next Step: Exponential Life, BBVA OpenMind.

[56] Angela E.B. Stewart, Hana Vrzakova, Chen Sun, Jade Yonehiro, Cathlyn Adele Stone, Nicholas D. Duran, Valerie Shute, and Sidney K. D'Mello. 2019. I Say, You Say, We Say: Using Spoken Language to Model Socio-Cognitive Processes during Computer-Supported Collaborative Problem Solving. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 194 (nov 2019), 19 pages.

[57] Peter Stone, Gal A Kaminka, Sarit Kraus, and Jeffrey S Rosenschein. 2010. Ad hoc autonomous agent teams: Collaboration without pre-coordination. In *Twenty-Fourth AAAI Conference on Artificial Intelligence*.

[58] Stefan Stürmer and Mark Snyder. 2010. *The psychology of prosocial behavior: Group processes, intergroup relations, and helping*. Wiley-Blackwell.

[59] Katia Sycara and Gita Sukthankar. 2006. Literature review of teamwork models. *Robotics Institute, Carnegie Mellon University* 31 (2006), 31.

[60] Tuomas Takko, Kunal Bhattacharya, Daniel Monsivais, and Kimmo Kaski. 2021. Human-agent coordination in a group formation game. *Scientific Reports* 11, 1 (2021), 1–10.

[61] Robin Thompson. 1980. Maximum likelihood estimation of variance components. *Statistics: A Journal of Theoretical and Applied Statistics* 11, 4 (1980), 545–561.

[62] Güliz Tokadlı, Kaitlyn Ouverson, Chase Meusel, Austin Garcia, Stephen B Gilbert, and Michael C Dorneich. 2018. An Analysis of Video Games Using the Dimensions of Human-Agent Interaction. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 62. SAGE Publications Sage CA: Los Angeles, CA, 716–720.

[63] Robert L Trivers. 1971. The evolution of reciprocal altruism. *The Quarterly review of biology* 46, 1 (1971), 35–57.

[64] Stephanie Tulk, Ryon Cumings, Taha Zafar, and Eva Wiese. 2018. Better know who you are starving with: Judging humanness in a multiplayer videogame. In *Proceedings of the Technology, Mind, and Society*. 1–6.

[65] Tomer D Ullman, Chris L Baker, Owen Macindoe, Owain Evans, Noah D Goodman, and Joshua B Tenenbaum. 2009. Help or hinder: Bayesian models of social goal inference. In *Proceedings of the 22nd International Conference on Neural Information Processing Systems*. 1874–1882.

[66] Alexandra Vtyurina, Denis Savenkov, Eugene Agichtein, and Charles LA Clarke. 2017. Exploring conversational search with humans, assistants, and wizards. In *Proceedings of the 2017 chi conference extended abstracts on human factors in computing systems*. 2187–2193.

[67] Deepika Yadav, Prerna Malik, Kirti Dabas, and Pushpendra Singh. 2019. Feedpal: Understanding Opportunities for Chatbots in Breastfeeding Education of Women in India. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 170 (nov 2019), 30 pages. https://doi.org/10.1145/3359272

[68] Xingchen Zhou, Pei-Luen Patrick Rau, and Xueqian Liu. 2021. " Time to Take a Break" How Heavy Adult Gamers React to a Built-In Gaming Gradual Intervention System. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–30.

[69] Brian D Ziebart, Andrew L Maas, J Andrew Bagnell, Anind K Dey, et al. 2008. Maximum entropy inverse reinforcement learning.. In *Aaai*, Vol. 8. Chicago, IL, USA, 1433–1438.