Conversational Agents on Your Behalf: Opportunities and Challenges of Shared Autonomy in Voice Communication for Multitasking

Yi Fei Cheng
Sony Computer Science Laboratories, Inc.
Tokyo, Japan
Okinawa Institute of Science and Technology Graduate University
Okinawa, Japan
Human-Computer Interaction Institute
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA
yifeic2@andrew.cmu.edu

Hirokazu Shirado
Human-Computer Interaction Institute
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA
shirado@cmu.edu

Shunichi Kasahara
Sony Computer Science Laboratories, Inc.
Tokyo, Japan
Okinawa Institute of Science and Technology Graduate University
Okinawa, Japan
kasahara@csl.sony.co.jp

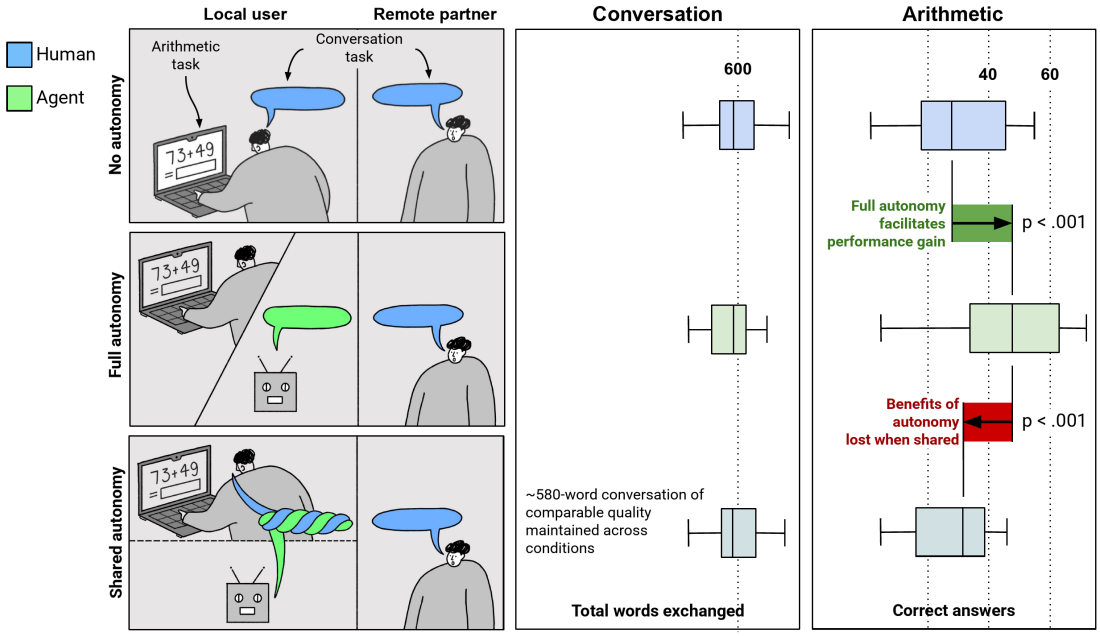

Figure 1: We present the results of a study with 18 dyads ($N = 36$) investigating the effect of sharing autonomy with a conversational agent on multitasking. Participant pairs played the roles of a local *user* and a remote *partner*. The user was tasked with conversing with the partner while performing arithmetic operations. They performed this multitasking with no support (*no autonomy*), by delegating the conversation entirely to an agent (*full autonomy*), or by *sharing autonomy* over the conversation. In all three conditions, users maintained a ~580-word conversation of comparable quality. Fully delegating the conversation to an agent enabled users to concentrate on their parallel task, leading to performance gains; however, when autonomy is shared, these benefits are reduced.

CHI '25, Yokohama, Japan

ACM ISBN 979-8-4007-1394-1/25/04
https://doi.org/10.1145/3706598.3714017

Abstract

Advancements in computational agents will enable them to act as surrogates for users in online communication, promising enhanced productivity by supporting multitasking. This capability may be especially powerful when combined with human control, allowing users to retain agency while achieving better performance

than either human or agent alone. However, it remains unclear how people might leverage this technology to multitask effectively. We present a study with 18 dyads exploring how users employ automated responses to support an arithmetic task while staying engaged in a voice call. Participants multitasked with a conversational agent under three levels of autonomy: none, shared, and full. Our findings indicate that fully automated systems can maintain conversational engagement, enabling users to multitask effectively. Surprisingly, shared autonomy hindered this ability. Based on our results, we discuss implications for designing shared autonomy in conversations, highlighting new considerations and challenges.

CCS Concepts

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

Keywords

computer-mediated communication, agents

ACM Reference Format:

Yi Fei Cheng, Hirokazu Shirado, and Shunichi Kasahara. 2025. Conversational Agents on Your Behalf: Opportunities and Challenges of Shared Autonomy in Voice Communication for Multitasking. In *CHI Conference on Human Factors in Computing Systems (CHI '25), April 26–May 01, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3706598.3714017

1 Introduction

Computational agents increasingly demonstrate the ability to model human behavior [60], including that of specific individuals [72]. Ever-growing advancements in deep-learning technologies, such as speech and video synthesis [16, 76], enable us to incorporate these agents into not only text-based interactive systems but also real-time multimodal communication media, such as voice or video meetings [41]. These agents may support users by serving as their surrogates, allowing users to engage in multiple tasks simultaneously, such as catching up on work while delegating a lightweight routine meeting for the agent to attend on their behalf [63].

However, it remains unclear how people will perceive and engage with these agents in communication settings. While surrogate agents assist users by interacting with others on their behalf, users may face new challenges related to delegation and supervision [25]. Balancing trust in an agent's autonomy with oversight can create cognitive strain. Finding the right level of intervention may also require extra effort. Therefore, the productivity benefits promised by computational agents must be carefully weighed against the potential cognitive burdens they impose.

Our study aimed to investigate the following research question: **how does the ability to offload conversational engagement to a computational agent influence user's multitasking abilities during voice communication?** To explore this, we prototyped a real-time voice communication system that enabled users to delegate conversational engagement to an agent. We evaluated the system through an experiment involving 18 pairs of participants (N=36), manipulating the agent's levels of autonomy. Each pair participated in a call about a pre-defined topic. One participant was assigned a secondary task, which they completed alone (*no autonomy*), with a conversational agent managing communication independently (*full autonomy*), or while sharing control (*shared autonomy*).

Our findings indicated that granting the system full autonomy to manage conversations improved users' multitasking performance, but it also caused discomfort as users worried about losing control. In contrast, allowing users to oversee and control the surrogate agent led to a loss of performance gains, as they felt the need to monitor the conversation closely to ensure alignment with their own responses, which negated the intended cognitive load reduction.

On the other hand, from the perspective of users' conversation partners, the use of the agent, whether in full or shared autonomy, did not significantly compromise conversation quality. Additionally, while users felt uneasy about offloading conversational tasks, their partners adapted well to any system imperfections.

In the remainder of the paper, we first introduce the concept of shared autonomy within a conversation in greater detail. For replicability, we then provide a detailed description of the experimental system we implemented to study multitasking with varying levels of shared autonomy. Next, we report on our experiment design and results. We conclude with a discussion of the opportunities and challenges of deploying computational agents as surrogates in real-time communication, highlighting implications for the design of future conversational systems.

2 Related Work

2.1 Shared Autonomy in Communication

There is a significant prior research the benefits of interactive systems where users delegate some autonomy to computational agents, instead of solely relying on direct control or full automation. In the design of user interfaces, such so-called *mixed-initiative* approaches [28] are said to effectively merge the strengths of direct manipulation and automated reasoning. In robotics, *shared autonomy* in teleoperation has been shown to support more efficient task completion by allowing users to assert goals while benefiting from automated control in complex tasks [57, 65]. Recently termed *human-{machine, agent, AI}* teaming, interaction, or collaboration [2, 15, 17], these paradigms demonstrate capabilities beyond what humans or computation can achieve alone, impacting areas from accessibility [24, 42] to creativity [73, 75].

In computer-mediated communication (CMC), the integration of computational agents has increasingly transformed platforms from simple transmission mediums to systems that can modify, augment, or generate communication content, enabling users to similarly share or delegate some level of autonomy to a computational entity during interpersonal interactions [25]. In the text domain, computational agents have optimized messages for better communication outcomes, such as conveying high status [64] or trustworthiness [50]. Kim et al. [37] used text recommendations to increase affectionate communication between romantic partners, while Argyle et al. [3] demonstrated the effectiveness of an AI assistant in promoting democratic discourse.

Beyond text-based communication, prior work has explored various visual and auditory manipulations, such as increasing one's attractiveness [44] or making a speaker sound more calm or authoritative [38]. Recent learning-based approaches have further enabled

the generation of entirely synthetic communication content for a specific individual, including speech segments [72] and facial expressions [36]. Coupled with the capabilities of language models to replicate believable human behavior [60], these technologies can facilitate shared autonomy in real-time conversations, allowing users to speak or delegate speaking to an agent as needed.

The ability of a conversational agent to speak for users can be adapted for similar communication goals as previous work. Additionally, there is interest in enabling users to clone themselves and parallelize tasks to boost productivity [63]. However, many open questions remain about how people may leverage and interact with this technology to optimize their communications broadly [25], including for the purpose of multitasking. Our work explores how users employ automated responses to support an arithmetic task while remaining engaged in a voice call. Although the literature on shared autonomy indicates that combining manual and automated control can enhance performance, our findings highlight challenges that hinder this in communication tasks.

2.2 Multitasking

Multitasking refers to carrying out two or more tasks simultaneously [71], such as responding to emails while attending a meeting. There is a substantial body of work on its prevalence [14], how it is prompted [20, 34, 52], and its effects on performance and well-being [33]. Generally, prior studies suggest that multitasking adversely affects memory [7, 51], performance [56], and incurs psychological costs [4, 53]. However, it remains a common practice [14, 33, 71] and can sometimes even offer benefits, such as improving creativity [35] and decision-making [70].

Of particular relevance to our work is prior research on multitasking during meetings. In their large-scale analysis of multitasking behaviors, Cao et al. [14] found that knowledge workers often multitasked either to increase productivity or to relieve anxiety through breaks. They additionally observed that sometimes multitasking would be prompted by external distractions that were outside of the workers' control. In Iqbal et al.'s study [33], they observed that people who multitasked during meetings did so to interleave other important activities. While multitasking, they paid peripheral attention to the meeting and engaged only when relevant. Although multitasking has its problems, preventing it entirely is neither realistic nor productive [54]. Instead, there is value in acknowledging the inevitability of multitasking as a natural component of online meetings and designing solutions to support its benefits [69].

In our work, we build on this approach by exploring the effects of providing users with the ability to delegate their conversation to an agent. This should ideally allow the user to focus on their secondary task while keeping their conversation partner engaged; however, as our results suggest, enabling this interaction is far from straightforward.

3 Concept: Sharing Autonomy in a Conversation

During conversations, multitasking requires engaging with the partner while managing a secondary task. However, people often struggle to divide their attention effectively between multiple tasks [56]. A personalized conversational agent could help users offload conversational responsibilities, allowing them to focus on other tasks. With surrogate agents, users can stay "present" in conversations while catching up later through summaries [69]. However, this *fully autonomous* mode eliminates the user's ability to participate in the conversation, which may be undesirable [25].

An alternative approach is to support *shared autonomy* in the conversation. Rather than relinquishing complete control, users can choose when to engage directly and when to delegate responses to the agent. Users can then remain involved in the conversation while focusing on secondary tasks when necessary, potentially enhancing multitasking performance beyond what is achievable with either full automation or no automation at all.

We explore these potentials of computational agents in our experimental conversation system by controlling agent autonomy at three levels: *no autonomy* (no agents), *full autonomy*, and *shared autonomy* (see the details in Section 4.5).

4 Methods: Enabling Shared Autonomy in a Conversation

We operationalized the concept of conversational agents with three levels of autonomy in an experimental communication system. Our design was guided by three core objectives: **(O1)** to enable seamless interleaving of agent-driven and user responses while maintaining a consistent speaker identity; **(O2)** to naturally engage the conversation partner; and **(O3)** to prioritize user input. We prioritized user input based on the assumption that even with advanced algorithms, the user is still best positioned to direct the conversation, consistent with previous work on shared autonomy [65].

To achieve these objectives, we developed a communication system in which an agent used pre-recorded segments of the user's speech, automatically extracted from a similar prior conversation (i. e., with a different user), to respond to a remote partner (Figure 2). We opted for pre-recorded speech segments instead of a voice cloning or conversion-based pipeline (e. g., ElevenLabs [1]) because, through early prototypes, we found that these alternatives resulted in noticeable discrepancies, which disrupted the perception of a consistent speaker identity **(O1)**.

4.1 Communication Channel

The system user and remote conversation partner's devices were connected via Zoom [79] to enable audio communication. All processing was implemented via a Python wrapper around the Zoom client of the system user, which operated on the application's inputs and outputs through virtual audio cables [12].

4.2 Conversational Agent

Our conversational agent predicted and injected pre-recorded *responses* and *backchannels* from a user into the conversation with a remote conversation partner. *Responses* refer to longer statements people typically give during a conversation, occupying a turn in the dialogue. *Backchannels* are brief vocalizations such as 'hmm' or 'uh-huh' [77]. These backchannels represent a basic form of human conversations [27] and were included to demonstrate additional engagement and understanding **(O2)** [21, 40]. Our conversational

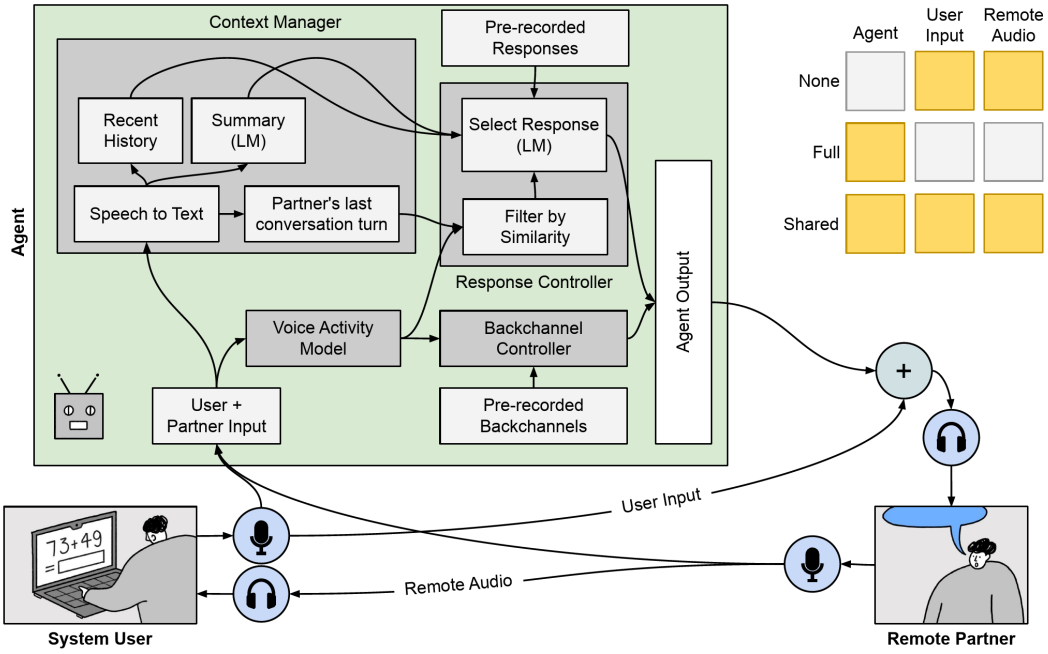

Figure 2: Overview of the experimental *shared autonomy* system. The system enables real-time communication between a local user and a remote partner. Users can speak directly to their partner or delegate the conversation to an agent. The table in the top right indicates the features enabled for each level of autonomy (highlighted in yellow).

agent consisted of four core components: a *Voice Activity Model*, *Context Manager*, *Response Controller*, and *Backchannel Controller*.

4.2.1 Voice Activity Model. Introducing responses and backchannels (*O2*) requires accurate modeling of interpersonal conversation dynamics. This includes predicting when the partner is about to yield their turn (i. e., turn-shifts) and determining appropriate moments for backchannels. To achieve this, our agent leverages Ekstedt and Skantze's Voice Activity Projection (VAP) model [22, 31, 32], an open-source, real-time voice activity detection model that has demonstrated state-of-the-art performance.

Given audio inputs from the user and the remote partner, VAP computes two key values for our system: (1) the probability that the remote partner will yield their conversation turn (i. e., *predicted turn-shifts*) and (2) the probability of an upcoming user backchannel (i. e., the *predicted backchannel*). The specific moment of a *turn-shift event* is determined by applying a threshold to the predicted turn-shift probabilities, while *backchannel events* is determined using z-score-based peak signal detection algorithm [9]. These events govern the timing of automated responses and backchannels.

4.2.2 Context Manager. Beyond determining the appropriate timing, response contents must also be contextually relevant to build on the ongoing conversation (*O2*). To capture this context, our agent transcribes the dialogue in real-time, using Google's Speech-to-Text API [48], and maintains a *recent history* of the last five conversation turns. In addition, the agent leverages a language model (Open AI's GPT-4o [59], temperature = 0.5; prompt details in Appendix A) to continuously update a paragraph-long *summary* of the conversation's content with each transcribed utterance.

4.2.3 Response Controller. The agent's *Response Controller* component manages both the timing and selection of responses. Appropriate response times are identified as a parameterized period (ii. e., *response delay time*) following a *turn-shift event* during which the partner has not spoken.

For response selection, the agent adopts a method similar to Zulfikar et al. [80]. For each newly transcribed utterance from the remote partner, we first filter potential responses from our pre-recorded set based on relevance and then use a Language Model to identify the single most pertinent response.

To filter for potential responses, the agent computes a vectorized embedding of the content of the remote partner's most recent conversation turn. This embedding is used to search for the most semantically similar responses by comparing them to the embeddings of the response entries in the pre-recorded set, which has been precomputed. The comparison is performed using a nearest neighbor search based on the cosine similarity between the vectors [39]. All embeddings are calculated using a pre-trained all-MiniLM-L6-v2 sentence transformer model [66], which maps input text to a 384-dimensional vector. In this study, the agent filters for the 10 most relevant responses.

To select a single response, the agent processes the filtered results using a second Language Model instance (GPT-4, temperature = 0.5; prompts are detailed in Appendix A). This model is prompted to choose a response based on the *summary* and *recent history*, following guidelines to ensure responses naturally continue the conversation while avoiding repetition and revisiting discussion

points already covered. To minimize the risk of introducing inappropriate responses, the model is instructed to refrain from selecting a response if none are deemed suitable.

4.2.4 Backchannel Controller. The agent's *Backchannel Controller* component determined the timing and selection of backchannels. Backchannel are introduced at *backchannel events*, with their frequency moderated by a *minimum backchannel time* parameter. Backchannels are selected at random since they do not carry significant meaning.

4.3 User Input and Interface

Users can speak directly to their remote partner through their computer, as they would in a natural conversation or voice call (*O3*). The system actively monitors for the user's voice and, if detected, interrupts and prevents any concurrent input from the agent to prioritize the user's input. This allows users to correct the direction of the discourse as needed. User inputs are further processed by the *Context Manager* to ensure consistency in subsequent responses.

Following Son et al. [69], our system includes a basic front-end interface that presents real-time dialogue transcripts. The interface also notifies users through a background color change when the agent was unable to generate an appropriate response, alerting them to the need for input.

4.4 Response and Backchannel Extraction

Our agent uses a pre-recorded set of responses and backchannels to communicate with human counterpart. To facilitate their collection, we implemented an *automatic extraction* feature that, given a recorded dialogue, extracts relevant responses and backchannels. For response extraction, the system generates a speaker-diarized transcription of the dialogue using Google Speech-to-Text [48], and segments the recording based on conversation turns. Each user speaking turn is recorded as a separate entry in the database. For backchannels extraction, the system uses the VAP model [22] to compute predicted backchannels and applies a peak detection algorithm [68, 74] to identify one-second windows that may correspond to moments when the user was providing backchannel feedback.

4.5 Levels of Autonomy

Our experimental system is adapted to support the three levels of autonomy described in Section 3 as follows:

No autonomy: Conversing with an agent that operates with no autonomy is achieved by disabling the conversation agent (Section 4.2) in our system, reducing it to a normal voice call where the user interacts directly with their partner.

Full autonomy: Conversing with an agent that operates with full autonomy is achieved by disabling user input and their access to the interface (Section 4.3). This precludes their ability to intervene in or monitor the conversation. In this case, the dialogue is managed entirely by the agent.

Shared autonomy: Conversing while sharing autonomy with an agent is enabled by supporting both user input (Section 4.2) and agent contributions (Section 4.2). This setup allows the user to listen to and participate in the conversation as they naturally would in a voice call, while delegating the conversation to the agent when they are not actively engaged, as in full autonomy. We include an example snippet of shared autonomy conversations from our experiment (section 5) in Appendix B to illustrate the interaction.

5 Experiment

We conducted an experiment with 18 participant pairs (N=36). Participants took on the roles of a local *user* and a remote *partner*. Each pair was tasked with discussing fictional student profiles (*conversation task*), while the *user* participant was also asked to perform math calculations as a secondary task (*arithmetic task*). Each session consisted of three conversation segments, each focused on a different student profile, with the *user* multitasking alongside a conversation agent operating under a different LEVEL OF AUTONOMY (no autonomy, full autonomy, and shared autonomy). Knowledge of multitasking and agent usage was withheld from *partner* until the end. *Users* were also instructed to multitask without disclosing their engagement with the secondary task to simulate a realistic meeting scenario [14]. The study was approved by the local IRB.

5.1 Tasks

5.1.1 Conversation Task. We designed a conversation task where the user would engage in an open-ended but partially repetitive conversation, simulating scenarios where automating parts of the conversation would be feasible and beneficial. The task centered around the evaluation of a fictional student profile. Participants received different information sets for the task. *Users* was given evaluation standards, including criteria and probing questions about the student's motivation, research interests, academic background, and personal qualities. Meanwhile, *partners* received 2-page student profiles formatted like a statement of purpose.

Pairs discussed a single profile in each conversation, with the goal of helping *user* participants gather enough information for an evaluation. No decisions were made during the conversations; the focus was on information exchange. The task positioned participants in asymmetric roles: *users* primarily asked questions based on evaluation criteria, while *partners* provided responses and additional details. This setup allowed the *user* to guide the discussion with standard probing questions, enabling structured interactions that could incorporate automated conversational elements. Each segment lasted five minutes.

5.1.2 Arithmetic Task. To simulate a multitasking scenario, only *user* participants were also assigned a secondary task involving arithmetic operations (specifically 2-digit addition and subtraction). Users were instructed to perform these operations as quickly and accurately as possible, with the incentive that if they solved 70 questions within the 5-minute segment, both they and their partner would receive a monetary bonus. However, they were instructed to do so discreetly, ensuring that their multitasking remained undetected by their partner throughout the conversation segment.

We chose arithmetic operations as the secondary task to represent a cognitively demanding activity, following prior multitasking studies [10, 19, 67]. To control task difficulty, questions were presented in randomized blocks of four: one addition and one subtraction without carry-over, and one addition and one subtraction with carry-over.

5.2 Conditions

Our experiment evaluates participants' multitasking performance across three LEVEL OF AUTONOMY conditions: multitasking with a conversation agent that operated with *no autonomy*, *full autonomy*, and *shared autonomy*. Each condition was implemented using our experimental system, as detailed in Section 4, with specific adaptations for each condition outlined in Section 4.5. The order of conditions was counterbalanced using a Latin Square design.

5.3 Apparatus

Participants performed tasks in two adjacent experimental rooms, where no sound could be heard between them to simulate a remote conversation setting. They were equipped with an Audio-Technica ATH-102USB headset, serving as both audio input and output on each end of our system. The experiment ran on a Dell G15 5520 for the system user and a Dell Precision 5570 for the conversation partner, with both connected via the university network. A detailed illustration of the experimental setup is provided in Appendix E.

5.4 Procedure

Figure 3 provides an overview of the study procedure. The entire procedure took approximately 120 minutes per session.

(1) **Introduction (5-10 min)**. After providing consent and demographic information, participants were introduced to the study as an investigation of communication behaviors. They were informed that they would participate in three five-minute discussions about different fictional student profiles (Section 5.1.1). Multitasking and agent usage details were initially withheld. Participants were assigned roles (*user* or *remote partner*) based on preference or randomly.
(2) **Task preparations: partner (45-60 min)**. *Partner* participants prepared in a separate room, familiarizing themselves with three student profiles. Their objective was to provide sufficient information on each profile to assist in the evaluation. They were guided on key aspects, such as academic background, research interests, and personal qualities, and encouraged to spend approximately 15 minutes per profile.
(3) **Task preparations: user (45-60 min)**. *User* participants were briefed on their objectives of multitasking with a conversation agent. To collect material for this agent, users first engaged in an 8-minute training session with the experimenter, where they practiced the primary conversation task. This training session was recorded to capture the participants' responses and backchannels (see Section 4.4). Following this, participants took part in an additional training session with the experimenter to become familiar with the various levels of autonomy of the conversational agent. During these sessions, system parameters were adjusted to align with each participant's preferences, such as response delays and backchannel settings.
(4) **Tasks (45 min)**. *User* and *partner* participants were paired for three conversation segments under different autonomy levels (Section 5.2). Each segment focused on a separate student profile. After each segment, participants completed questionnaires on their experiences with the conversation and arithmetic tasks.
(5) **Post-Study Questionnaire (5 min)**. Participants completed a post-study questionnaire comparing their preferences across conditions. *Partner* participants were then briefed on the study's purpose, including the use of conversational agents, and asked to identify conditions retrospectively.
(6) **Exit Interview (15 min)**. In a semi-structured group interview, participants discussed their impressions of the system, preferred conditions, appropriate contexts for the system, potential improvements, and any suspicions about automation. The condition order was revealed during this session (see Appendix D).

5.5 Measures

We evaluated participants' performance and perceptions through a range of quantitative and self-reported metrics, as described below. The full questionnaires are reported in Appendix C.

5.5.1 Conversation Quality and Behavior. We assessed conversation quality using subjective evaluations adapted from prior studies [8, 13, 29, 49, 55]. *User* participants rated their presence (*user presence (self)*), their partner's presence (*user presence (other)*), attention to their partner (*user attention (self)*), their partner's attention (*user attention (other)*), and their own naturalness (*user naturalness*). In the *full autonomy* condition, these ratings reflected users' expectations of their agent's performance. Additionally, users reported their *conversation task load* via the Raw NASA-TLX questionnaire [26]. *Partner* participants rated mutual presence and attention (*partner presence (self)*, *partner presence (other)*, *partner attention (self)*, *partner attention (other)*), the user's naturalness (*partner naturalness*), and the interactivity of the conversation (*partner interactivity*). After all segments, *user* participants ranked the segments by their perceived success (*conversation success*), while *partner* participants ranked them by their engagement (*engagement*). We also analyzed conversation behavior, including the number of *total words*, *user words*, *partner words*, and *conversation turn shifts*.

5.5.2 Arithmetic Task Performance. We evaluated performance on the secondary arithmetic task by analyzing the number of *correct answers*, *accuracy* (percentage of correct answers), and *standard deviation in response times* to measure performance consistency across conditions. Additionally, *user* participants reported their subjective *arithmetic task load* using the RTLX questionnaire[26] for each segment and ranked the segments by their perceived success in the arithmetic task (*arithmetic success*).

5.5.3 Partner Perception of Automation. Knowledge of the usage of an automated response system was withheld from the partner until the end of the study, including the order of the conditions. After disclosure, we asked if they had suspected its use (*suspected autoresponse*) and to identify the correspondence between conversation segments and conditions (*session condition*).

5.5.4 System Usage. To understand system usage, we measured the agent's *responses* (number and total words spoken), *backchannels*, and *user overrides* (instances where users interrupted the agent). Participants also evaluated the usability of the *full* and *shared autonomy* conditions using the System Usability Scale [11].

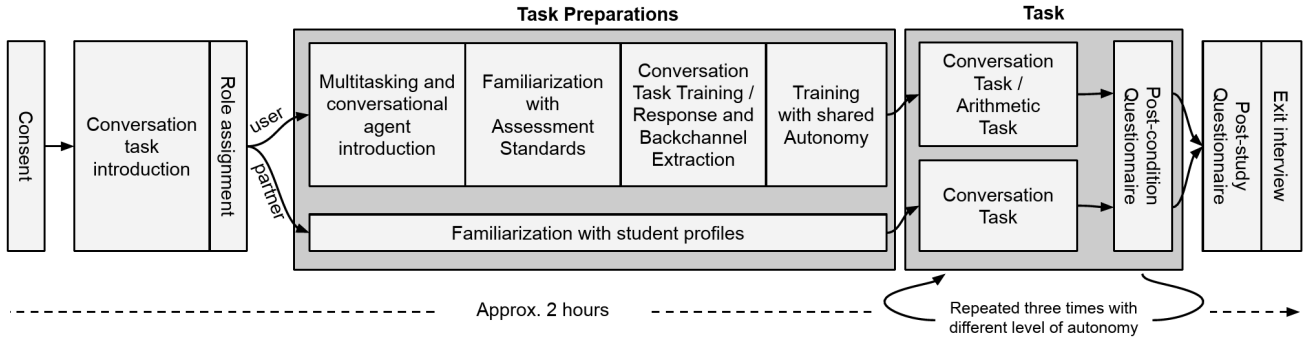

Figure 3: Overview of our study procedure.

5.6 Participants

Our power analysis determined that 15 dyads were needed to detect a medium effect size (0.5) with 95% power and an α of 0.05. We also drew referenced prior multitasking studies (e. g., [58]).

Following the sample size analysis, we recruited 36 participants in 18 dyads (18 male, 18 female) between the age of 20 and 54 ($M = 28$, $SD = 7$) from a university through message groups, social networks, publicly distributed posters, and word-of-mouth. Participants represented 16 countries and reported conversational proficiency in English. Regarding audio and video communication tools, 16 participants reported using them daily, 7 used them 2-3 times a week, 9 used them once a week, and 4 used them at least once a month. Familiarity between participants within each dyad varied: 8 pairs interacted almost daily, 1 pair weekly, 2 pairs monthly, and 7 pairs had never interacted before. Nevertheless, performance analyses showed no meaningful differences in their performance based on the system and pair familiarity. Participants received a base compoensation of ¥3000 (approx. $21) and a performance bonus of ¥500 (approx. $4) for the arithmetic task.

6 Results

We evaluated the performance of the conversation and arithmetic tasks across the three LEVELS OF AUTONOMY, analyzed the partner's perception of automation (including suspicion and condition identification), and assessed the agent's contributions in the *full* and *shared autonomy* conditions (e. g., responses, words, backchannels).

For interval data (e.g., words exchanged, turn shifts, correct answers), we used repeated measures ANOVA with LEVEL OF AUTONOMY (*none*, *full*, *shared*) as a within-subject variable. Violations of normality (Shapiro-Wilk $p < .05$) or ordinal data were addressed with the Friedman test. When equal variance assumptions were violated (Mauchley's test $p < .05$), Greenhouse-Geisser correction was applied. Pairwise comparisons with Bonferroni adjustments were conducted as needed. All statistical analyses were performed using IBM SPSS 29 [30].

6.1 Conversation Quality and Behavior

Figure 4 summarizes the results on the effect of LEVEL OF AUTONOMY on conversation performance. We calculated composite scores for perceived conversation quality. For *partners*, the score averaged six metrics: *partner presence (self)*, *presence (other)*, *attention (self)*, *attention (other)*, *naturalness*, and *interactivity*, showing strong reliability (Cronbach's $\alpha = 0.88$). For *users*, the score averaged five metrics: *user presence (self)*, *presence (other)*, *attention (self)*, *attention (other)*, and *naturalness*, also demonstrating high reliability (Cronbach's $\alpha = 0.81$).

We first discuss the *partner* participants' subjective evaluations of the conversations. There was no significant effect of LEVEL OF AUTONOMY on *partner perceived conversation quality* ($\chi^2(2) = 2.94$, $p = .23$) or *engagement* rankings ($\chi^2(2) = .73$, $p = .70$). These results suggest that **partners consistently perceived the conversation quality as high across all LEVELS OF AUTONOMY**. Even when the agent fully took over communication, *partners* rated the conversations similarly to those where *users* communicated directly or assisted the agent.

On the other hand, *user*'s subjective ratings of conversations significantly differed across conditions. Pairwise comparisons revealed that *users* rated conversation quality 1.4x higher in *no autonomy* ($p < .001$) and 1.3x higher in *shared autonomy* ($p = .006$) than in *full autonomy*. Users also ranked their *conversation success* higher in *no autonomy* than in *full autonomy* ($p = .012$). *Users* reported significantly different *conversation task loads* across conditions ($\chi^2(2) = 12.55$, $p = .002$), experiencing 34% lower task load in *full autonomy* than in *no autonomy* ($p = .021$) and 26% lower in *full autonomy* than in *shared autonomy* ($p = .015$). These results suggest that **fully delegating conversations to the agent reduced user task load but lowered their perceived conversation quality and success**. This contrasts with *partners*, who rated conversation quality consistently high across all levels of autonomy.

We observed similar trends in participants' conversation behavior. There were no significant differences in *total words* exchanged or *turn shifts* across the LEVEL OF AUTONOMY ($\chi^2(2) = 3$, $p = .23$ and $F_{1,2} = 2.00$, $p = .15$, $\eta_p^2 = 0.11$, respectively). However, the number of words contributed by *partner* participants ($F_{1,2} = 4.12$, $p = .025$, $\eta_p^2 = 0.20$) and *user* participants ($\chi^2(2) = 32.44$, $p < .001$) differed significantly. Pairwise tests showed that *partners* spoke, on average, 50 fewer words (about 10% of $M = 445$, $SD = 89$ words per conversation) in *no autonomy* compared to *full autonomy* ($p = 0.04$). For *users*, significant differences were found between all conditions ($p < .01$). As expected, *user* participants contributed no words in *full autonomy*, and 42% fewer words in *shared autonomy* compared

to *no autonomy*. These results indicate that **a similar quantity of conversation was generally maintained across the LEVELS OF AUTONOMY**. However, *partners* spoke slightly less in *full autonomy* than in *no autonomy*, suggesting areas for improving the agent's ability to engage human partners effectively.

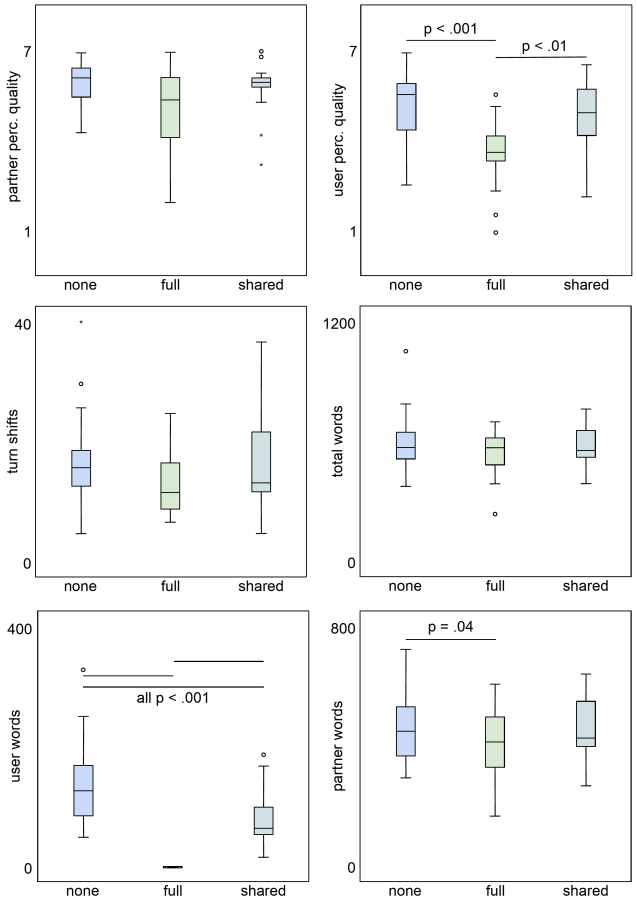

Figure 4: Effect of LEVEL OF AUTONOMY on *partner perceived conversation quality, user perceived conversation quality, turn shifts, total words, user words,* and *partner words.*

6.2 Arithmetic Task Performance

We analyzed *users'* arithmetic task performance to understand how well they managed multitasking during conversations and how agents influenced performance across autonomy levels (Figure 5). We found a significant effect of LEVEL OF AUTONOMY on the number of *correct answers* ($F_{1,2} = 30.13$, $p < .001$, $\eta_p^2 = 0.64$; ANOVA). *Users* answered 33% fewer questions in *no autonomy* and 36% fewer in *shared autonomy* compared to *full autonomy* (both $p < .001$). However, there were no significant differences between conditions in terms of *accuracy*, *response time standard deviation*, *arithmetic task load*, or *arithmetic success*. These findings confirm that **fully delegating the conversation to an agent allowed users to focus entirely on the arithmetic task**, which is unsurprising. However, in the *shared autonomy* condition, users did not see any performance gains despite offloading parts of the conversation. This suggests that ***shared autonomy* may introduce complexities that prevent users from fully utilizing the time saved by automation**. We explore factors influencing this outcome in Sections 7 and 8.

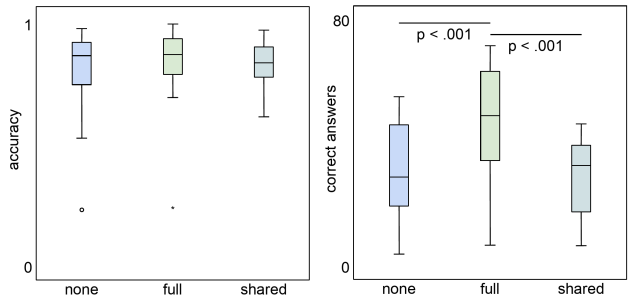

Figure 5: Effect of LEVEL OF AUTONOMY on *accuracy* and number of *correct answers* in arithmetic task.

6.3 Partner Perception of Automation

Most *partner* participants did not suspect the use of automation during the conversations and struggled to identify the LEVEL OF AUTONOMY even after being informed about the agent's involvement. **Only 3 out of 18 partners suspected automation during the conversations**. After being briefed (without details on specific conditions), they correctly identified *no autonomy* in 8 out of 18 cases, *full autonomy* in 10 cases, and *shared autonomy* in only 5 cases (Figure 6). These results suggest that **relatively engaging conversations with a consistent human-speaker identity were maintained across all conditions**.

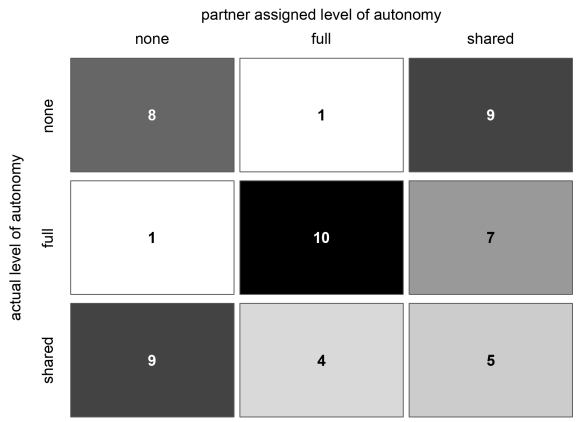

Figure 6: Confusion matrix for partner assignments of LEVEL OF AUTONOMY.

6.4 Agent Contributions

Figure 7 summarizes the system usage in the *full* and *shared autonomy* conditions. In the *full autonomy* condition, the agent contributed an average of 7 *responses* ($SD = 3$), which amounted to

131 (SD = 69) *words*, and 9 *backchannels* (SD = 6). In the *shared autonomy* condition, the agent contributed an average of 3 *responses* (SD = 2), which amounted to 51 (SD = 38) *words*. On average, this represented approximately 40% of the total words contributed by the user (M = 80, SD = 49 words) and agent combined. Agent responses were rarely interrupted or overridden by the user (M = 0.5, SD = 1.0 interruptions). The agent also contributed 7 *backchannels* (SD = 6) on the user's behalf. Users rated full automation with a usability score of 75 and shared autonomy with 69, both in the "good" range [6]. These results suggest that the agent's contributions in *full autonomy* matched the user's in *no autonomy* and helped share conversation responsibility in *shared autonomy*.

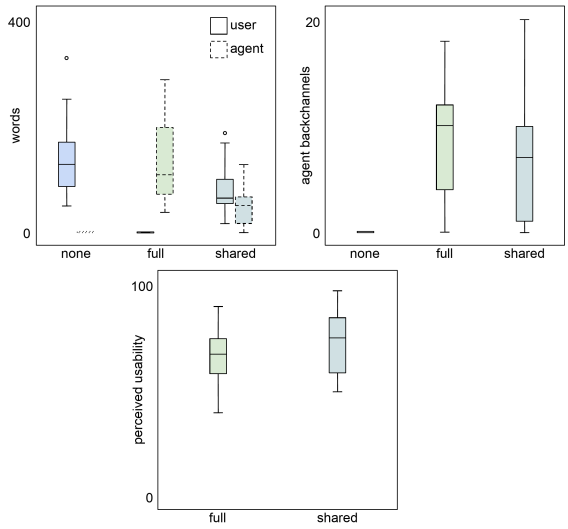

Figure 7: System usage metrics across *levels of autonomy*, including words contributed by the user and agent, agent backchannels, and perceived usability.

7 Qualitative findings

We examined the interview data to gain further insight into the experience of conversing with an agent that operates at varying LEVELS OF AUTONOMY. The interviews were recorded across all sessions and transcribed. A bottom-up approach to thematic analysis was taken to analyze the transcripts. The researcher who conducted the studies manually generated over 450 open codes and organized them into themes. These themes were iterated with input from other members of the research team. Below, we present the final set of themes, reflecting the user and partner's perspectives on interacting with a conversation surrogate and usage considerations.

7.1 System User Perspectives

In the full autonomy condition, users reported the lowest conversation task load but also expected the lowest conversation quality. Interviews confirmed that while **full autonomy helped users focus on their secondary task, it also caused discomfort due to their inability to engage in the conversation** (N=9). P3 shared, "I was more focused on my task [with full autonomy], but I was also more worried about what was going on." Some participants likened the experience to being "in the dark" (P5). This discomfort was often linked to mistrust in the agent's abilities (N=7), with P1 remarking, "I was kind of worried you were thinking, what is that dumb*** talking about?"

In contrast, users in the shared autonomy condition felt **more control while offloading parts of the conversation** (N=6). P11 appreciated the ability to intervene, saying, "I stepped in to guide [the conversation] in a specific direction." P7 mentioned relying on the agent to cover gaps when they were "focused on math" and "forgot to ask [P8] some questions."

However, many users (N=13) found that **shared autonomy increased the effort required to participate in conversations**. Some (N=4) felt distracted by auditory feedback, while most (N=11) felt compelled to closely monitor the dialogue. Concerns about the agent's ability to represent them accurately (N=13) and occasional issues like slow or out-of-context responses (N=7, N=8) contributed to this need for oversight.

Even when the system responded appropriately, users still felt the need to monitor it (N=9). P3 speculated, "Maybe it's because, like, now I'm not very confident with the system." P30 suggested that monitoring might persist until the system could replicate the user's thought process entirely. Similarly, P29 noted feeling "pre-programmed to anticipate" their own responses and compare them to the agent's.

Overall, the results indicate that **access to the conversation in shared autonomy inadvertently increased user attention to the dialogue**, reducing the multitasking benefits of automation.

7.2 Remote Partner Perspectives

In 15 out of 18 sessions, remote partners (*partner* participants) did not suspect any use of automation. Of the three sessions where automation was suspected, two partners cited prior knowledge of the research team's work, and one session revealed automation due to a system glitch, where the agent repeatedly provided irrelevant responses.

After learning about the use of automation, all remote partners noted peculiarities that, in hindsight, indicated automation. The most common issue was receiving out-of-context responses (N=12), including overly generic replies (N=5) and responses that ignored prior dialogue (N=10), such as repeating questions or changing topics abruptly. For instance, P6 recalled being asked "Can you elaborate?" on simple sentences. Similarly, P20 noted how the student was repeatedly misgendered in the full autonomy conversation: "I was using 'they' for all of them, and ... no matter how much I said it, you always used 'he.' "

Partners also attributed automation to differences in the user's behavior (N=10), such as missing characteristic humor. For instance, P10 pointed out how the user they were paired with tended to crack jokes, but this behavior was missing in one of the sessions. Interestingly, more positively perceived interactions were also considered signs of automation (N=4), for instance, if the system user was "being too nice" (P30), "helpful" (P22), or "fluent" (P8). Some also suspected automation when conversational patterns seemed atypical (N=4) For instance, P36 noted, "Usually there's more maybe pleasantries in a conversation, before you get into, like, the actual

question." Similarly, P24 attributed their suspicion of automation to the lack of guidance when they were lost: "I felt a human would have stopped me and said, 'All right, tell me about this.' "

However, **the behaviors remote partners cited as automated were not always actually automated** (N=6). For instance, while P2 attributed repeated questions to the presence of automation, it was a human error. As P1 clarified, "I didn't actually hear it. That wasn't the automated response." Similarly, before being disclosed about automation, **many participants had interpreted the slightly awkward automated behaviors as human** (N=8). For instance, when encountering a repeated question, some partners assumed it was a request for clarification (N=4). As P16 explained, "I was like, maybe I wasn't clear, so let me clarify." Sometimes, when given a slightly out-of-context response, partners assumed their partner was trying to move the conversation forward (N=2). As P24 remarked, "Maybe he got enough information about one topic ... it was not unnatural for him to jump from place to place."

In 8 of 18 sessions, the interview conversations explicitly revealed **an asymmetry in how users and partners perceived the interaction**. For example, while the response timing might feel uncomfortable for the system user, it was perceived as natural by the remote partner (N=4). Similarly, while the automated backchanneling behavior was sometimes perceived as unnatural by the user, it was overlooked by the partner (N=2).

7.3 Usage & Applications

Overall, the concept of conversational surrogates elicited a range of responses. While some participants found the idea beneficial (e. g., P32: "I would use it, like, constantly, every day, in every meeting"), others expressed significant reservations about its use (e. g., P11: "For me, personally and professionally, neither would it be appropriate to use"). The perceived utility and acceptability of a conversation surrogate depended on several dimensions.

A core consideration was whether the conversation served a **transactional or social purpose** (N = 9). According to P28, they would not use it with their romantic partner because "I have conversations with them to... communicate and build more bond, rather than just extracting information." In conversations with friends and family, participants generally did not see a purpose in substituting themselves with a "soulless" agent (P11) that can neither "understand nor empathize" (P8), preferring to participate actively or not at all. Relating to this, many participants expressed concerns that their use of a surrogate might be perceived as **disrespectful** (N = 15). Emphasizing a **social expectation of reciprocal engagement**, P5 stated, "If I have a conversation with someone, I want that person ... to pay attention."

Participants generally favored **less consequential use cases** (N = 12). For instance, P16 envisioned using a surrogate only in conversations that "don't affect [their] life and [their] future." Participants also leaned towards **simpler interactions** (N = 9), such as in group conversations where they listen passively (P13). Several participants saw potential in agents managing more "routine" (P5) and structured conversations and answering simple questions. In contrast, participants were more skeptical about using agents in more interactive open discussions.

8 Discussion

In this paper, we presented an empirical study with 18 dyads investigating how conversing with an agent operating at varying LEVELS OF AUTONOMY affects multitasking performance. In the following, we discuss our main results and their implications for future implementations of conversational surrogates.

8.1 Effect of LEVEL OF AUTONOMY on Multitasking

8.1.1 Full Autonomy Maintained Conversation Engagement. Our results show that the delegation system we implemented effectively engaged with human partners during the five-minute conversation task. In all conversation segments, including full autonomy, most partner participants did not realize automated responses were used and found the conversation quality similar to those with direct user participation (Figure 4). That said, occasional imperfections arose, such as responses that did not align with the conversational context. Interestingly, human partners often adapted to these errors by interpreting them as their own mistakes or misunderstandings by the user (Section 7.2).

8.1.2 Full Autonomy Improved Secondary Task Performance. In addition to maintaining engagement, full autonomy allowed users to improve their secondary task performance (Figure 5). This suggests that people had difficulty allocating cognitive resources in our study task of juggling conversation and arithmetic. Delegating the conversation to an agent may help alleviate this challenge.

8.1.3 Asymmetries in Perception of Full Autonomy Conversations. While partners did not perceive differences in conversation quality between levels of autonomy, many users felt that the quality of the agent's conversations was significantly lower than their own (Figure 4). Users also expressed discomfort with fully delegating their conversations to the agent (see Section 7). This discomfort stemmed in part from their complete exclusion from the conversation, which many participants described as feeling "in the dark." This aligns with prior research on social support systems [23], which highlights the importance of visibility, awareness, and accountability in the delegation process. Participants' unfamiliarity with agent-assisted conversations further contributed to this discomfort. As all participants were using the delegation system for the first time, they may not have developed sufficient trust in the agent.

8.1.4 Visibility in Shared Autonomy: A Double-edged Sword. In contrast to the full autonomy condition, the shared autonomy condition provided greater visibility, awareness, and accountability within the delegation process. As a result, although users delegated about half of their contributions to the agent, they still reported a higher level of satisfaction with the quality of the conversations, rating it similar to their direct interactions (Figure 4).

However, the benefit of delegation systems for multitasking diminished under shared autonomy (Figure 5). When users could intervene in the agent's performance, their secondary task performance declined to levels similar to the condition without the agent. Although the shared autonomy system provided more visibility, it also imposed a cognitive burden comparable to not using the agent.

Our qualitative analysis suggests that this suppression effect may stem from an asymmetry in how users perceive the conversation. Users expressed discomfort with fully delegating the conversation to the agent. We speculate that this led them to continuously monitor the agent's discussions under shared autonomy even when they were not actively participating. This tendency was likely what undermined the utility of the delegation system.

While this increase in oversight efforts might be mitigated as users develop greater trust in conversational agents, it may also reflect a deeper, fundamental challenge rooted in human perceptions of self-identity. Indeed, our qualitative analysis suggests that users' intrinsic motivation to align the agent's behavior with their own identity and social norms also contributed to the suppression effect of shared autonomy. When computational agents act as a surrogate for a user, the agent's responses may inevitably draw the user's attention, as the interaction outcomes are perceived as extensions of their own behavior. This highlights that identity preservation and self-representation are critical considerations in designing shared autonomy systems.

8.2 Designing for Conversational Surrogates

8.2.1 Usage in Social versus Transactional Conversations. As reported in Section 7.3, there are various scenarios in which computational agents can enhance conversations, but also contexts in which their usage is deemed inappropriate or questionable. The value of these agents depends largely on the nature of the conversation, particularly whether it serves a social (or affectional) or transactional (or instrumental) purpose [18].

Our results suggest that within conversations aimed at building and maintaining social bonds, the use of conversational surrogates was generally less meaningful. Employing delegation systems in such contexts is seen as disrespectful. In addition, especially for interacting with friends or family, participants expressed a desire to engage in conversations for their own sake, highlighting that their personal presence carries intrinsic value.

Conversely, in conversations that are purely transactional — like when exchanging information relevant to tasks — using conversational surrogates is generally seen as more acceptable. However, this raises a question about the relevance of the user's identity in these interactions. For example, when making restaurant reservations, it is arguably not important whether the booking is made by the user themselves or by a Google Assistant [47].

However, in practice, the distinction between social and transactional interactions is often blurred and dynamic [5, 78], as most interactions serve a combination of both purposes [18]. We argue that this overlap presents an opportunity where conversational surrogates may be applied most effectively (Figure 8).

For example, in work-related meetings, conversations often serve both *instrumental* purposes, like keeping team members updated, and *social* purposes, such as fostering inter-collegial relationships. In these situations, if users are unable to participate due to other commitments, a surrogate could help them contribute to information exchange (transactional) while maintaining their presence in online meetings (social). As Leong et al. [43] recently demonstrated, designing an agent that visually and vocally represents the user — serving as a surrogate rather than merely acting on their behalf — is crucial for fostering feelings of presence and trust among collaborators.

Similarly, in instructional settings, these agents can help scale personalized engagement by addressing student questions individually and managing discussions for inclusivity (transactional) [46]. Acting as the instructor's surrogate, these agents can also foster deeper connections by embodying a respected authority while also strengthening student-instructor relationships (social) [61]. Prior work has shown that an instructor's identity can significantly influence student engagement and learning outcomes [62].

Recently, some political candidates have even used computational agents to disseminate their views (transactional), to enable more interactive and personalized engagement (social) at scale while maintaining message consistency [45]. We anticipate that similar applications of computational agents will emerge across various domains, scaling up "personal" interactions within one-to-many relationships.

8.2.2 Balancing Delegation and Control in Self-Representation. Given the interplay between social and transactional goals, our study suggests that designing conversational surrogates should navigate the trade-offs between delegation and control in self-representation. As discussed in Section 8.1, neither the full nor shared autonomy designs evaluated in our study managed the trade-off effectively. Full autonomy prioritized delegation at the cost of control. Our shared autonomy design provided more control with increased visibility, but this also impeded delegation.

As language technology advances, we expect the communication capability of computational agents to improve and enable longer and more complex conversations. However, we believe that this trade-off challenge is unlikely to be resolved solely by technical improvement. In our study, users expected agents not only to respond plausibly but also to accurately reflect how they themselves would respond. This means that surrogate systems may need to effectively model an individual's thought processes, rather than merely simulating general human conversations. To the authors, this prospect seems questionable. While we cannot dismiss this possibility, we suggest that addressing this trade-off through interaction design may provide a more promising direction for future work in the meantime.

8.2.3 Opportunities for Interaction Design. Our findings suggest that merely offering visibility and control may not facilitate the effective utilization of delegation systems. Alternative interaction methods might better address the balance between delegation and individual agency. For instance, enabling users to preview and queue the agent's responses or specify higher-level can help maintain control while reducing the need for constant oversight. An intriguing direction could involve developing abstractions that bridge the user's communication intent and the agent's word-by-word verbalization. Building on the work of Son et al. [69], new methods might improve users' understanding of conversation context for better re-engagement. However, introducing such additional mechanisms may add complexities, potentially increasing cognitive load. This highlights the need for careful evaluation to ensure that these designs truly enhance user experience without undermining the intended benefits of delegation systems.

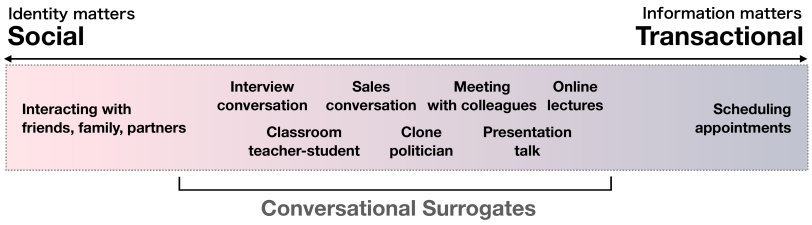

Figure 8: Conversations may be mapped to a transactional to social continuum. We argue the emergent use cases for conversational surrogates lie in the middle, representing interactions that serve both transactional and social needs.

8.3 Limitations and Future Work

Our study is subject to several limitations, which we discuss below.

8.3.1 Conversation Task Constraints. First, it should be noted that our study represents only a constrained conversational context regarding the stakes and interaction demands. In particular, our participants engaged in a discussion about fictional characters, for which they were not evaluated on the outcomes. The task design also resulted in a structured conversation, where the user primarily asked questions.

Interestingly, despite the low-stakes nature of the conversation task, our user participants still dedicated significant attention to monitoring the agent's responses. This potentially speaks to the strength of social considerations, as they seemed to manifest even in such an artificial and inconsequential setting. Based on this, we speculate that with higher stakes, the requirements for effective delegation and representation will become more demanding, but the exact effects of the task stakes should be evaluated further. For instance, future studies could consider augmenting our current task to provide incentives to get the student's evaluation right.

Future work should also evaluate the effects of surrogates on tasks with different interaction demands. We anticipate that our system and results may be directly applicable in conversations where the agent can simply listen passively and occasionally ask questions or in those that follow a standard structure, like semi-structured interviews. However, deploying agents in conversations that require active participation to inform decisions is more challenging. In these situations, the conversation may rely not just on static prior knowledge but also on evolving opinions. This means the surrogate user would need to dynamically express how they wish to be represented, which poses a greater interaction challenge.

Analyzing practical usage scenarios can provide valuable insights. For instance, some participants suggested using agents that passively "listen" to group conversations. However, when we examine the interactions that occur within a group conversation, it becomes clear that the verbal backchanneling mechanism currently implemented in our system may not be suitable. Brief affirmations like "I see" or "uh-huh" could disrupt the flow of conversation in these settings. Instead, visual cues, such as nodding, are more commonly used to demonstrate engagement. This observation indicates a potential need for visual backchanneling in future systems.

Similarly, in online learning scenarios (see Section 8.2.1), the goal of enabling scalable personalized instruction is both promising and challenging. For instance, agents designed to provide personalized answers to students during lectures may face complex questions that exceed the system's knowledge base. These edge cases highlight the need for mechanisms that allow for escalation to human instructors. Additionally, we need to take into account the effort required by instructors to implement these agents. In our study, we developed the agents' responses based on previous conversations. In an educational context, this approach could involve using data from earlier course iterations, such as past lecture recordings. For example, an agent in a computer science course could be pre-programmed with solutions to common programming errors identified in previous semesters. Investigating how historical data can facilitate agent development and alleviate instructor workload could be a valuable direction for future research.

8.3.2 Secondary Task. In our study, we used an arithmetic task based on the prior multitasking literature [10, 19, 67]. Performing the arithmetic task while maintaining a conversation turned out to be detrimental, but still manageable to some extent. Hence, the overhead of managing the use of an agent outweighed the benefits of delegation. However, in more demanding tasks, such as those that are more linguistically dependent or have higher memory requirements, the benefits of delegation may be more pronounced. Future studies should examine the trade-off between delegation and representation with respect to a wider range of secondary tasks.

8.3.3 Sample Size & Generalizability. For our study, we recruited 18 dyads from a university context. Although we believe that this sample size is sufficient for an initial investigation, replicating the study with a larger and more diverse participant pool will be valuable to further strengthen the generalizability of our findings.

8.3.4 Novelty Effects & Discomfort. Since none of our participants had experience with having an agent speak on their behalf, it is likely that the results exhibited some level of novelty effects. We observed that these novelty effects manifested as some level of discomfort in both the full and shared autonomy conditions. In the latter, user discomfort may have contributed to their monitoring behavior, which ultimately diminished their multitasking abilities. While greater familiarity could improve usage, it is uncertain if it fully bridges the performance gap. In addition, perspectives and behaviors around their use may change as social norms evolve, especially as the use of conversational agents becomes more prevalent. Therefore, future longitudinal studies and evaluations that control for familiarity and comfort can provide valuable insights.

8.3.5 Agent Imperfections. While the agent we built arguably managed the conversation well, it was not perfect, as indicated by the

slightly reduced participation on the part of the partner and qualitative reports of quirks. Our results suggest that the use of shared autonomy is not just a matter of technical sophistication but also a human-factor limitation. That said, we believe there is value in periodically reevaluating our results with novel technologies.

9 Conclusions

In this work, we presented the results of a controlled study with 18 pairs of participants investigating how varying levels of autonomy in conversational agents affect users' multitasking performance. Our results indicate that fully delegating autonomy to an agent enables improved multitasking performance but is discomforting for users, as it requires them to entirely relinquish their control. While sharing autonomy reduces this discomfort, it inadvertently triggers a need to monitor the conversation, thereby diminishing the benefits of delegating conversational responsibility for multitasking. By highlighting these challenges in sharing autonomy, our work provides valuable insights for the future design of agent-mediated real-time communication systems.

Acknowledgments

We thank all involved peers, participants, and anonymous reviewers, especially Elijah Claggett, David Lindlbauer, Pedro Lopes for their input throughout the project. We also give a special thanks to Shannon Yeung for her lovely illustrations and support. Yi Fei Cheng was supported by the Croucher Foundation. This work was supported by JST Grant Number JPMJPR23I4, JPMJPF2205 and JST Moonshot R&D Program Grant Number JPMJMS2013. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the sponsors.

References

[1] [n. d.]. 11EleventLabs. https://elevenlabs.io/. Accessed: 2024-09-04.

[2] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for Human-AI Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300233

[3] Lisa P. Argyle, Christopher A. Bail, Ethan C. Busby, Joshua R. Gubler, Thomas Howe, Christopher Rytting, Taylor Sorensen, and David Wingate. 2023. Leveraging AI for democratic discourse: Chat interventions can improve online political conversations at scale. *Proceedings of the National Academy of Sciences* 120, 41 (2023), e2311627120. https://doi.org/10.1073/pnas.2311627120 arXiv:https://www.pnas.org/doi/pdf/10.1073/pnas.2311627120

[4] Brian P. Bailey and Joseph A. Konstan. 2006. On the need for attention-aware systems: Measuring effects of interruption on task performance, error rate, and affective state. *Computers in Human Behavior* 22, 4 (2006), 685–708. https://doi.org/10.1016/j.chb.2005.12.009 Attention aware systems.

[5] Shreya Bali, Pranav Khadpe, Geoff Kaufman, and Chinmay Kulkarni. 2023. Nooks: Social Spaces to Lower Hesitations in Interacting with New People at Work. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 614, 18 pages. https://doi.org/10.1145/3544548.3580796

[6] Aaron Bangor, Philip Kortum, and James Miller. 2009. Determining what individual SUS scores mean: Adding an adjective rating scale. *Journal of usability studies* 4, 3 (2009), 114–123.

[7] Louise Barkhuus. 2005. "Bring your own laptop unless you want to follow the lecture": alternative communication in the classroom. In *Proceedings of the 2005 ACM International Conference on Supporting Group Work* (Sanibel Island, Florida, USA) *(GROUP '05)*. Association for Computing Machinery, New York, NY, USA, 140–143. https://doi.org/10.1145/1099203.1099230

[8] Frank Biocca, Chad Harms, and Jenn Gregg. 2001. The networked minds measure of social presence: Pilot test of the factor structure and concurrent validity. In *4th annual international workshop on presence, Philadelphia, PA*. 1–9.

[9] J.P.G van Brakel. 2014. Robust peak detection algorithm using z-scores. https://stackoverflow.com/questions/22583391/peak-signal-detection-in-realtime-timeseries-data. https://stackoverflow.com/questions/22583391/peak-signal-detection-in-realtime-timeseries-data

[10] Valdimar Briem and Leif R Hedman. 1995. Behavioural effects of mobile telephone use during simulated driving. *Ergonomics* 38, 12 (1995), 2536–2562.

[11] J Brooke. 1996. SUS: A quick and dirty usability scale. *Usability Evaluation in Industry* (1996).

[12] Vincent Burel. 2024. VB-CABLE Virtual Audio Device. https://vb-audio.com/Cable/. Accessed: 2024-09-04.

[13] Runze Cai, Nuwan Nanayakkarawasam Peru Kandage Janaka, Shengdong Zhao, and Minghui Sun. 2023. ParaGlassMenu: Towards Social-Friendly Subtle Interactions in Conversations. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 721, 21 pages. https://doi.org/10.1145/3544548.3581065

[14] Hancheng Cao, Chia-Jung Lee, Shamsi Iqbal, Mary Czerwinski, Priscilla N Y Wong, Sean Rintel, Brent Hecht, Jaime Teevan, and Longqi Yang. 2021. Large Scale Analysis of Multitasking Behavior During Remote Meetings. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 448, 13 pages. https://doi.org/10.1145/3411764.3445243

[15] Tara Capel and Margot Brereton. 2023. What is Human-Centered about Human-Centered AI? A Map of the Research Landscape. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 359, 23 pages. https://doi.org/10.1145/3544548.3580959

[16] Sam W. T. Chan, Tamil Selvan Gunasekaran, Yun Suen Pai, Haimo Zhang, and Suranga Nanayakkara. 2021. KinVoices: Using Voices of Friends and Family in Voice Interfaces. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 446 (oct 2021), 25 pages. https://doi.org/10.1145/3479590

[17] Nazli Cila. 2022. Designing Human-Agent Collaborations: Commitment, responsiveness, and support. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 420, 18 pages. https://doi.org/10.1145/3491102.3517500

[18] Leigh Clark, Nadia Pantidi, Orla Cooney, Philip Doyle, Diego Garaialde, Justin Edwards, Brendan Spillane, Emer Gilmartin, Christine Murad, Cosmin Munteanu, Vincent Wade, and Benjamin R. Cowan. 2019. What Makes a Good Conversation? Challenges in Designing Truly Conversational Agents. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300705

[19] Mary Czerwinski, Edward Cutrell, and Eric Horvitz. [n. d.]. Instant messaging and interruption: Influence of task type on performance. Citeseer.

[20] Laura Dabbish, Gloria Mark, and Víctor M. González. 2011. Why do i keep interrupting myself? environment, habit and self-interruption. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vancouver, BC, Canada) *(CHI '11)*. Association for Computing Machinery, New York, NY, USA, 3127–3130. https://doi.org/10.1145/1978942.1979405

[21] Zijian Ding, Jiawen Kang, Tinky Oi Ting HO, Ka Ho Wong, Helene H Fung, Helen Meng, and Xiaojuan Ma. 2022. TalkTive: A Conversational Agent Using Backchannels to Engage Older Adults in Neurocognitive Disorders Screening. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 304, 19 pages. https://doi.org/10.1145/3491102.3502005

[22] Erik Ekstedt and Gabriel Skantze. 2022. Voice Activity Projection: Self-supervised Learning of Turn-taking Events. In *Proc. Interspeech 2022*. 5190–5194. https://doi.org/10.21437/Interspeech.2022-10955

[23] Thomas Erickson and Wendy A. Kellogg. 2000. Social translucence: an approach to designing systems that support social processes. *ACM Trans. Comput.-Hum. Interact.* 7, 1 (March 2000), 59–83. https://doi.org/10.1145/344949.345004

[24] Anhong Guo, Junhan Kong, Michael Rivera, Frank F. Xu, and Jeffrey P. Bigham. 2019. StateLens: A Reverse Engineering Solution for Making Existing Dynamic Touchscreens Accessible. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology* (New Orleans, LA, USA) *(UIST '19)*. Association for Computing Machinery, New York, NY, USA, 371–385. https://doi.org/10.1145/3332165.3347873

[25] Jeffrey T Hancock, Mor Naaman, and Karen Levy. 2020. AI-Mediated Communication: Definition, Research Agenda, and Ethical Considerations. *Journal of Computer-Mediated Communication* 25, 1 (01 2020), 89–100. https://doi.org/10.1093/jcmc/zmz022 arXiv:https://academic.oup.com/jcmc/article-pdf/25/1/89/32961176/zmz022.pdf

[26] SG Hart. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Human mental workload/Elsevier* (1988).

[27] Bettina Heinz. 2003. Backchannel responses as strategic responses in bilingual speakers' conversations. *Journal of Pragmatics* 35, 7 (2003), 1113–1142. https://doi.org/10.1016/S0378-2166(02)00190-X

[28] Eric Horvitz. 1999. Principles of mixed-initiative user interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Pittsburgh, Pennsylvania, USA) *(CHI '99)*. Association for Computing Machinery, New York, NY, USA, 159–166. https://doi.org/10.1145/302979.303030

[29] Erzhen Hu, Md Aashikur Rahman Azim, and Seongkook Heo. 2022. FluidMeet: Enabling Frictionless Transitions Between In-Group, Between-Group, and Private Conversations During Virtual Breakout Meetings. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 511, 17 pages. https://doi.org/10.1145/3491102.3517558

[30] IBM. 2024. IBM SPSS Software. https://www.ibm.com/spss. Accessed: 2024-05-03.

[31] Koji Inoue, Bing'er Jiang, Erik Ekstedt, Tatsuya Kawahara, and Gabriel Skantze. 2024. Multilingual Turn-taking Prediction Using Voice Activity Projection. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (Eds.). ELRA and ICCL, Torino, Italia, 11873–11883. https://aclanthology.org/2024.lrec-main.1036

[32] Koji Inoue, Bing'er Jiang, Erik Ekstedt, Tatsuya Kawahara, and Gabriel Skantze. 2024. Real-time and Continuous Turn-taking Prediction Using Voice Activity Projection. arXiv:2401.04868 [cs.CL] https://arxiv.org/abs/2401.04868

[33] Shamsi T. Iqbal, Jonathan Grudin, and Eric Horvitz. 2011. Peripheral computing during presentations: perspectives on costs and preferences. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vancouver, BC, Canada) *(CHI '11)*. Association for Computing Machinery, New York, NY, USA, 891–894. https://doi.org/10.1145/1978942.1979073

[34] Shamsi T. Iqbal and Eric Horvitz. 2007. Disruption and recovery of computing tasks: field study, analysis, and directions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '07)*. Association for Computing Machinery, New York, NY, USA, 677–686. https://doi.org/10.1145/1240624.1240730

[35] Chaitali Kapadia and Shimul Melwani. 2021. More tasks, more ideas: The positive spillover effects of multitasking on subsequent creativity. *Journal of Applied Psychology* 106, 4 (2021), 542.

[36] Shunichi Kasahara, Nanako Kumasaki, and Kye Shimizu. 2024. Investigating the impact of motion visual synchrony on self face recognition using real time morphing. *Scientific Reports* 14, 1 (2024), 13090.

[37] Taewook Kim, Jung Soo Lee, Zhenhui Peng, and Xiaojuan Ma. 2019. Love in Lyrics: An Exploration of Supporting Textual Manifestation of Affection in Social Messaging. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 79 (nov 2019), 27 pages. https://doi.org/10.1145/3359181

[38] Casey A. Klofstad, Rindy C. Anderson, and Susan Peters. 2012. Sounds like a winner: voice pitch influences perception of leadership capacity in both men and women. *Proceedings of the Royal Society B: Biological Sciences* 279, 1738 (2012), 2698–2704. https://doi.org/10.1098/rspb.2012.0311 arXiv:https://royalsocietypublishing.org/doi/pdf/10.1098/rspb.2012.0311

[39] Eyal Kushilevitz, Rafail Ostrovsky, and Yuval Rabani. 1998. Efficient search for approximate nearest neighbor in high dimensional spaces. In *Proceedings of the Thirtieth Annual ACM Symposium on Theory of Computing* (Dallas, Texas, USA) *(STOC '98)*. Association for Computing Machinery, New York, NY, USA, 614–623. https://doi.org/10.1145/276698.276877

[40] Divesh Lala, Pierrick Milhorat, Koji Inoue, Masanari Ishida, Katsuya Takanashi, and Tatsuya Kawahara. 2017. Attentive listening system with backchanneling, response generation and flexible turn-taking. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, Kristiina Jokinen, Manfred Stede, David DeVault, and Annie Louis (Eds.). Association for Computational Linguistics, Saarbrücken, Germany, 127–136. https://doi.org/10.18653/v1/W17-5516

[41] Patrick Yung Kang Lee, Ning F. Ma, Ig-Jae Kim, and Dongwook Yoon. 2023. Speculating on Risks of AI Clones to Selfhood and Relationships: Doppelganger-phobia, Identity Fragmentation, and Living Memories. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW1, Article 91 (apr 2023), 28 pages. https://doi.org/10.1145/3579524

[42] Sooyeon Lee, Rui Yu, Jingyi Xie, Syed Masum Billah, and John M. Carroll. 2022. Opportunities for Human-AI Collaboration in Remote Sighted Assistance. In *Proceedings of the 27th International Conference on Intelligent User Interfaces* (Helsinki, Finland) *(IUI '22)*. Association for Computing Machinery, New York, NY, USA, 63–78. https://doi.org/10.1145/3490099.3511113

[43] Joanne Leong, John Tang, Edward Cutrell, Sasa Junuzovic, Gregory Paul Baribault, and Kori Inkpen. 2024. Dittos: Personalized, Embodied Agents That Participate in Meetings When You Are Unavailable. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW2, Article 494 (Nov. 2024), 28 pages. https://doi.org/10.1145/3687033

[44] Tommer Leyvand, Daniel Cohen-Or, Gideon Dror, and Dani Lischinski. 2008. Data-driven enhancement of facial attractiveness. In *ACM SIGGRAPH 2008 Papers* (Los Angeles, California) *(SIGGRAPH '08)*. Association for Computing Machinery, New York, NY, USA, Article 38, 9 pages. https://doi.org/10.1145/1399504.1360637

[45] Gideon Lichfield. 2024. Meet your AI politician of the future. https://futurepolis.substack.com/p/meet-your-ai-politician-of-the-future. Accessed: 2024-12-05.

[46] Ziyi Liu, Zhengzhe Zhu, Lijun Zhu, Enze Jiang, Xiyun Hu, Kylie A Peppler, and Karthik Ramani. 2024. ClassMeta: Designing Interactive Virtual Classmate to Promote VR Classroom Participation. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 659, 17 pages. https://doi.org/10.1145/3613904.3642947

[47] Google LLC. 2024. Google Assistant, your own personal Google. https://assistant.google.com/. Accessed: 2024-12-09.

[48] Google LLC. 2024. Turn speech into text using Google AI. https://cloud.google.com/speech-to-text. Accessed: 2024-09-04.

[49] Matthew Lombard, Theresa B Ditton, and Lisa Weinstein. 2009. Measuring presence: the temple presence inventory. In *Proceedings of the 12th annual international workshop on presence*. International Society for Presence Research Los Angeles, CA, 1–15.

[50] Xiao Ma, Jeffrey T. Hancock, Kenneth Lim Mingjie, and Mor Naaman. 2017. Self-Disclosure and Perceived Trustworthiness of Airbnb Host Profiles. 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, 2397–2409. https://doi.org/10.1145/2998181.2998269

[51] Kevin P Madore, Anna M Khazenzon, Cameron W Backes, Jiefeng Jiang, Melina R Uncapher, Anthony M Norcia, and Anthony D Wagner. 2020. Memory failure predicted by attention lapsing and media multitasking. *Nature* 587, 7832 (2020), 87–91.

[52] Gloria Mark, Victor M. Gonzalez, and Justin Harris. 2005. No task left behind? examining the nature of fragmented work. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Portland, Oregon, USA) *(CHI '05)*. Association for Computing Machinery, New York, NY, USA, 321–330. https://doi.org/10.1145/1054972.1055017

[53] Gloria Mark, Yiran Wang, and Melissa Niiya. 2014. Stress and multitasking in everyday college life: an empirical study of online activity. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Toronto, Ontario, Canada) *(CHI '14)*. Association for Computing Machinery, New York, NY, USA, 41–50. https://doi.org/10.1145/2556288.2557361

[54] Sahar Mavali, Dongwook Yoon, Luanne Sinnamon, and Sidney S Fels. 2024. Time-Turner: A Bichronous Learning Environment to Support Positive In-class Multitasking of Online Learners. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 571, 15 pages. https://doi.org/10.1145/3613904.3641985

[55] Gerard McAtamney and Caroline Parker. 2006. An examination of the effects of a wearable display on informal face-to-face communication. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Montréal, Québec, Canada) *(CHI '06)*. Association for Computing Machinery, New York, NY, USA, 45–54. https://doi.org/10.1145/1124772.1124780

[56] Christopher A. Monk, Deborah A. Boehm-Davis, and J. Gregory Trafton. 2002. The Attentional Costs of Interrupting Task Performance at Various Stages. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 46, 22 (2002), 1824–1828. https://doi.org/10.1177/154193120204602210 arXiv:https://doi.org/10.1177/154193120204602210

[57] Stefanos Nikolaidis, Yu Xiang Zhu, David Hsu, and Siddhartha Srinivasa. 2017. Human-Robot Mutual Adaptation in Shared Autonomy. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction* (Vienna, Austria) *(HRI '17)*. Association for Computing Machinery, New York, NY, USA, 294–302. https://doi.org/10.1145/2909824.3020252

[58] Romain Nith, Yun Ho, and Pedro Lopes. 2024. SplitBody: Reducing Mental Workload while Multitasking via Muscle Stimulation. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 81, 11 pages. https://doi.org/10.1145/3613904.3642629

[59] OpenAI. 2024. Hello gpt-4o. https://openai.com/index/hello-gpt-4o/. Accessed: 2024-09-04.

[60] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative Agents: Interactive Simulacra of Human Behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (San Francisco, CA, USA) *(UIST '23)*. Association for Computing Machinery, New York, NY, USA, Article 2, 22 pages. https://doi.org/10.1145/3586183.3606763

[61] Pat Pataranutaporn, Valdemar Danry, Joanne Leong, Parinya Punpongsanon, Dan Novy, Pattie Maes, and Misha Sra. 2021. AI-generated characters for supporting personalized learning and well-being. *Nature Machine Intelligence* 3, 12 (2021), 1013–1022.

[62] Pat Pataranutaporn, Joanne Leong, Valdemar Danry, Alyssa P. Lawson, Pattie Maes, and Misha Sra. 2022. AI-Generated Virtual Instructors Based on Liked or Admired People Can Improve Motivation and Foster Positive Emotions for Learning. In *2022 IEEE Frontiers in Education Conference (FIE)*. 1–9. https://doi.

org/10.1109/FIE56618.2022.9962478
[63] Nilay Patel. 2024. The CEO of Zoom wants AI clones in meetings. https://www.theverge.com/2024/6/3/24168733/zoom-ceo-ai-clones-digital-twins-videoconferencing-decoder-interview. Accessed: 2024-09-04.
[64] Ellie Pavlick and Joel Tetreault. 2016. An Empirical Analysis of Formality in Online Communication. *Transactions of the Association for Computational Linguistics* 4 (03 2016), 61–74. https://doi.org/10.1162/tacl_a_00083 arXiv:https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00083/1567360/tacl_a_00083.pdf
[65] Siddharth Reddy, Anca D. Dragan, and Sergey Levine. 2018. Shared Autonomy via Deep Reinforcement Learning. arXiv:1802.01744 [cs.LG] https://arxiv.org/abs/1802.01744
[66] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, Hong Kong, China, 3982–3992. https://doi.org/10.18653/v1/D19-1410
[67] Joshua S Rubinstein, David E Meyer, and Jeffrey E Evans. 2001. Executive control of cognitive processes in task switching. *Journal of experimental psychology: human perception and performance* 27, 4 (2001), 763.
[68] scipy. 2024. Peak Finding. https://github.com/scipy/scipy/blob/v1.14.1/scipy/signal/_peak_finding.py#L729-L1010 GitHub repository.
[69] Seoyun Son, Junyoug Choi, Sunjae Lee, Jean Y Song, and Insik Shin. 2023. It is Okay to be Distracted: How Real-time Transcriptions Facilitate Online Meeting with Distraction. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 64, 19 pages. https://doi.org/10.1145/3544548.3580742
[70] Cheri Speier, Joseph S Valacich, and Iris Vessey. 1999. The influence of task interruption on individual decision making: An information overload perspective. *Decision sciences* 30, 2 (1999), 337–360.
[71] Minhyang (Mia) Suh, Frank Bentley, and Danielle Lottridge. 2018. "It's Kind of Boring Looking at Just the Face": How Teens Multitask During Mobile Videochat. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 167 (nov 2018), 23 pages. https://doi.org/10.1145/3274436
[72] Supasorn Suwajanakorn, Steven M. Seitz, and Ira Kemelmacher-Shlizerman. 2017. Synthesizing Obama: learning lip sync from audio. *ACM Trans. Graph.* 36, 4, Article 95 (jul 2017), 13 pages. https://doi.org/10.1145/3072959.3073640
[73] Craig Vear. 2021. Creative AI and Musicking Robots. *Frontiers in Robotics and AI* 8 (2021). https://doi.org/10.3389/frobt.2021.631752
[74] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods* 17 (2020), 261–272. https://doi.org/10.1038/s41592-019-0686-2
[75] Justin D. Weisz, Mary Lou Maher, Hendrik Strobelt, Lydia B. Chilton, David Bau, and Werner Geyer. 2022. HAI-GEN 2022: 3rd Workshop on Human-AI Co-Creation with Generative Models. In *Companion Proceedings of the 27th International Conference on Intelligent User Interfaces* (Helsinki, Finland) *(IUI '22 Companion)*. Association for Computing Machinery, New York, NY, USA, 4–6. https://doi.org/10.1145/3490100.3511166
[76] Mika Westerlund. 2019. The emergence of deepfake technology: A review. *Technology innovation management review* 9, 11 (2019).
[77] Sheida White. 1989. Backchannels across cultures: A study of Americans and Japanese. *Language in Society* 18, 1 (1989), 59–76. https://doi.org/10.1017/S0047404500013270
[78] Viviana A. Zelizer. 2000. The Purchase of Intimacy. *Law & Social Inquiry* 25, 3 (2000), 817–848. https://doi.org/10.1111/j.1747-4469.2000.tb00162.x
[79] Inc. Zoom Video Communications. 2024. Zoom. https://zoom.us/. Accessed: 2024-09-04.
[80] Wazeer Deen Zulfikar, Samantha Chan, and Pattie Maes. 2024. Memoro: Using Large Language Models to Realize a Concise Interface for Real-Time Memory Augmentation. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 450, 18 pages. https://doi.org/10.1145/3613904.3642450

A Prompt Templates

A.1 Summarize

A.1.1 System Prompt.

```
Update the CURRENT SUMMARY of an ongoing conversation
between 'USER' and 'PARTNER' based on the NEW
TRANSCRIPTS provided.

Guidelines:
- The summary should be concise, no longer than a
paragraph.
- Only summarize the contents of the discussion, not who
is speaking.
- Ensure the summary reflects any new information or
changes in the conversation.
- Return the updated summary.
```

A.1.2 User Prompt.

```
NEW TRANSCRIPT:
{transcript}

UPDATED SUMMARY:
```

A.2 Select Response

A.2.1 System Prompt.

```
Review the conversation SUMMARY and RECENT HISTORY
between 'YOU' and 'PARTNER'. From a list of POTENTIAL
RESPONSES, choose the one that best continues the
dialogue as 'YOU'.

GUIDELINES:
- Do not repeat yourself.
- Avoid revisiting points raised in the SUMMARY unless
necessary.
- Only return the numeric identifier of the selected
response.
- If a response would interrupt 'PARTNER', or if no
provided response is suitable, return 0.
- Be stringent in your selection; the chosen response
must fit seamlessly into the ongoing dialogue.
```

A.2.2 User Prompt.

```
SUMMARY:
{summary}

RECENT HISTORY:
{recent history}

POTENTIAL RESPONSES:
{responses}

RESPONSE:
```

B Shared Autonomy Conversation Example

```
PARTNER:
Yeah, they mostly talk about, like, these kind of
projects that they've worked on. So in terms of
```

```
extracurriculars, like, nothing about hobbies or
anything in here.

USER:
Okay. And what about teaching or mentoring
experience?

PARTNER:
Yes. So they don't actually mention that much mentoring
experience, even though one of their goals is to become
a professor. And I think maybe they haven't had
the chance because they're still in their bachelors, and
they're hoping to get that experience from their Ph.D.
years. It isn't written here, but it sounds like
one of the biggest opportunities that they want when
they're doing their Ph.D. is not just the research
and their increasing their own  understanding,
but also to be able to teach others, so  they
don't have teaching experience according to
this, right now.

AGENT (USER):
So, do you think they'd fit in well within... um...
within OIST... within that specific research unit?

PARTNER:
So I think it would be good, like if more questions
could be asked, for instance, like what you asked
about extracurriculars and even teaching experience.
From their application itself so they're very like
competent in terms of the work they've done. It's
excellent for someone that's just a bachelor
student. And yeah, I think like they make a good
fit, maybe like an interview with them personally
would help better. But yeah.

AGENT (USER):
So do you think they have the necessary skills and,
personality?

PARTNER:
Oh, yeah. So, not the personality part, I'm not
sure, but for sure their skills are up there and
their motivation is up there as well. So it was
a bit hard to gauge the personality from this.

USER:
What about their strengths and weaknesses?

PARTNER:
Hm, Okay. So their strengths actually...
```

C Questionnaire Items

C.1 Post-Condition Questionnaire (User)

Our post-condition questionnaire for the *user* participant included items evaluating task load (Q1-Q7; RTLX [26]) and conversation quality (Q8-Q12; based on [8, 13, 29, 49, 55]). All items used 7-point Likert scales.

- Q1 How mentally demanding was the {conversation, arithmetic} task? (1-Very Low, 7-Very High)
- Q2 How physically demanding was the {conversation, arithmetic} task? (1-Very Low, 7-Very High)
- Q3 How hurried or rushed (temporally demanding) was the pace of the {conversation, arithmetic} task? (1-Very Low, 7-Very High)
- Q4 How successful were you in completing the {conversation, arithmetic} task? (1-Failure, 7-Perfect)
- Q5 How hard did you have to work (effort) to accomplish your level of performance in the {conversation, arithmetic} task? (1-Very Low, 7-Very High)
- Q6 How insecure, discouraged, irritated, stressed, and annoyed (frustration) were you while completing the {conversation, arithmetic} task? (1-Very Low, 7-Very High)
- Q7 I often felt as if my partner and I were in the same voice call together. (1-Strongly Disagree, 7-Strongly Agree)
- Q8 I think my partner often felt as if we were in the same voice call together. (1-Strongly Disagree, 7-Strongly Agree)
- Q9 I paid close attention to my partner. (1-Strongly Disagree, 7-Strongly Agree)
- Q10 My partner paid close attention to me. (1-Strongly Disagree, 7-Strongly Agree)
- Q11 I acted naturally at all times during the conversation. (1-Strongly Disagree, 7-Strongly Agree)

For the *full* and *shared* autonomy conditions, *user* participants were additionally asked about their perceived usability of the conversation system (from the System Usability Scale [11]). All items used 5-point agreement Likert scales (1-Strongly Disagree, 5-Strongly Agree).

- Q12 I think that I would like to use this system frequently.
- Q13 I found the system unnecessarily complex.
- Q14 I thought the system was easy to use.
- Q15 I think that I would need the support of a technical person to be able to use this system.
- Q16 I found the various functions in this system were well integrated.
- Q17 I thought there was too much inconsistency in this system.
- Q18 I would imagine that most people would learn to use this system very quickly.
- Q19 I found the system very cumbersome to use.
- Q20 I felt very confident using the system.
- Q21 I needed to learn a lot of things before I could get going with this system.

C.2 Post-Condition Questionnaire (Partner)

Our post-condition questionnaire for the *user* participant included items and conversation quality (based on [8, 13, 29, 49, 55]). All items used 7-point Likert scales (1-Strongly Disagree, 7-Strongly Agree).

- Q1 I often felt as if my partner and I were in the same voice call together.
- Q2 I think my partner often felt as if we were in the same voice call together.

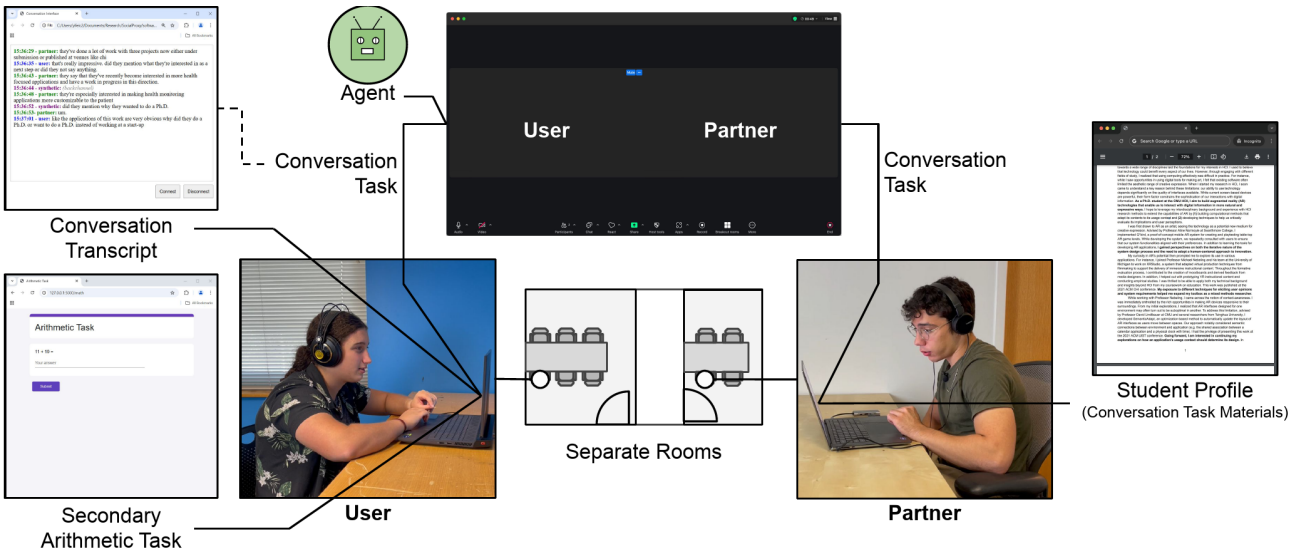

Figure 9: Experimental Setup. The user and partner participants complete their tasks in separate rooms to simulate a remote communication environment. The user participant is tasked with maintaining a conversation with their partner while simultaneously performing a secondary arithmetic task. In the shared autonomy condition, the user shares conversational responsibilities with an automated agent. Meanwhile, the partner participant focuses on reporting a student profile. All voice communication occurs through Zoom. The user completes the arithmetic task on their computer, which also displays a transcript of the conversation. The partner participant can access either a digital or physical copy of the student profile for reference during their conversation.

Q3 I paid close attention to my partner.
Q4 My partner paid close attention to me.
Q5 My partner acted naturally at all times during the conversation.
Q6 The conversation seemed highly interactive.

C.3 Post-Study Questionnaire (User)

Our post-study questionnaire for the *user* participant asked them to rank the conditions as follows.

Q1 Please rank the previous sessions in order of how successful you felt in completing the arithmetic task. (1-Most Successful, 3-Least Successful)
Q2 Please rank the previous sessions in order of how successful you felt in maintaining the conversation. (1-Most Successful, 3-Least Successful)

C.4 Post-Study Questionnaire (Partner)

Our post-study questionnaire for the *partner* participant was completed in two phases. First, participants ranked the conversation segments by engagement as follows.

Q1 Please rank the previous sessions in order of how engaged you felt in the conversation. (1-Most Engaged, 3-Least Engaged; segments labeled by number)

In the second phase, partner participants were first debriefed about the multitask and the usage of a conversational agent. They then answered the following questions.

Q2 Did you suspect an AI was involved during the task (i.e., before you were debriefed)? (Options: Yes, No)
Q3 Knowing that in some sessions the conversation was either fully or partially automated, which condition do you think was present in Session 1? (Options: Fully automated, Partially automated, Real)
Q3 Knowing that in some sessions the conversation was either fully or partially automated, which condition do you think was present in Session 2? (Options: Fully automated, Partially automated, Real)
Q3 Knowing that in some sessions the conversation was either fully or partially automated, which condition do you think was present in Session 3? (Options: Fully automated, Partially automated, Real)

D Exit Interview

Our exit interview consisted of the following questions:

Q1 What did you think about the experience overall?
Q2 Which condition did you prefer and why?
Q3 In what situations or applications would it be appropriate to use our system?
Q4 In what situations or applications would it be inappropriate?
Q5 How would you potentially modify the system?
Q6 What strategies did you use to multitask?
Q7 Did you suspect a conversational agent was involved? If not, how did you rationalize any of the differences you observed?
Q8 In retrospect, what behaviors give it away that a conversational agent was involved?

Q9 When did you let the agent speak for you? When did you deem it more appropriate to takeover?
Q10 Were there instances where the auto-response was incorrect? How did you handle or respond to these situations?

E Apparatus

We provide an illustration of our experimental setup in Figure 9.